UNITED STATES PATENT APPLICATION

FOR

REFERENCE CLONES AND SEQUENCES FOR NON-SUBTYPE B ISOLATES OF HUMAN IMMUNODEFICIENCY VIRUS TYPE 1

BY

BEATRICE H. HAHN, FENG GAO AND GEORGE M. SHAW

09144419

TITLE OF THE INVENTION

REFERENCE CLONES AND SEQUENCES FOR NON-SUBTYPE B ISOLATES OF HUMAN IMMUNODEFICIENCY VIRUS TYPE 1

This work was funded by grants RO1 AI25291; and NO1 AI35170 from the National Institutes of Health. Therefore, the government may have certain rights in the invention.

FIELD OF THE INVENTION

The present invention is in the field of virology. The invention relates to the nucleotide sequences of the genomes of 11 molecular clones for non-subtype B isolates of human immunodeficiency virus type 1 (HIV-1), and nucleic acids derived therefrom. This invention also relates to peptides encoded by and/or derived from the nucleic acid sequences of these molecular clones, and host cells containing these nucleic acid sequences and peptides. The invention also relates to diagnostic methods, kits and immunogens which employ the nucleic acids, peptides and/or host cells of the invention.

BACKGROUND OF THE INVENTION

A critical question facing current AIDS vaccine development efforts is to what extent HIV-1 genetic variation has to be considered in the design of candidate vaccines (11,21,42,72). Phylogenetic analyses of globally circulating viral strains have identified two distinct groups of HIV-1, a major M group and an O group (33,45,61,62). Within the M group, ten sequence subtypes (A-J) have been proposed (29,30,45,72). Sequence variation among viruses belonging to these different lineages is extensive, with envelope amino acid sequence variation ranging from 24% between different subtypes to 47% between the two different groups. Given this extent of diversity, the question has been raised whether immunogens based on a single virus strain can be expected to elicit immune responses effective against a broad spectrum of viruses, or whether vaccine preparations should include mixtures of genetically divergent antigens and/or be tailored toward locally circulating strains (11, 21, 42, 72). This is of particular concern in developing countries where multiple subtypes of HIV-1 are known to co-circulate and where subtype B viruses, which have been the source

2

for most current candidate vaccine preparations (10, 21), are rare or nonexistent (5, 24, 40, 72).

Although the extent of global HIV-1 variation is well defined, little is known about the biological consequences of this genetic diversity and its impact on cellular and humoral immune responses in the infected host. In particular, it remains unknown whether subtype specific differences in virus biology exist that need to be considered for vaccine design. Only a comprehensive analysis of genetically defined representatives of the various groups and subtypes will address the question of whether certain variants differ in fundamental viral properties and whether such differences will need to be incorporated into vaccine strategies. Obviously, such studies require well-characterized reference reagents, in particular full length and replication competent molecular clones that can be used for functional and biological studies.

Full-length reference sequences representing the various subtypes are also urgently needed for phylogenetic comparisons. Until about 1994, it was generally thought that individuals do not become infected with multiple distinct HIV-1 strains, and so the possibility that recombination between divergent viruses could contribute to the evolution of HIV-1 was not widely considered. However, recent analyses of subgenomic (23,52,54,58) as well as full-length HIV-1 sequences (7,18,53,60) identified a surprising number of HIV-1 strains which clustered in different subtypes in different parts of their genome. All of these originated from geographic regions where multiple subtypes co-circulated and are the results of co-infections with highly divergent viruses (52,60,62).

Recombinant viruses can be detected because their phylogenetic affinities vary depending on the region of genome analyzed. A useful initial approach is to examine the extent of sequence divergence/similarity between a new sequence and a bank of reference sequences of different subtypes, for example as a diversity plot (18), or using the RIP program (75); if the extent of relative similarity to different subtypes varies along the sequence, this may indicate that the sequence is a recombinant. However, fuller investigation must involve a phylogenetic approach, comparing trees derived by analyses of different regions of the genome, and assessing the confidence of phylogenetic clustering by a statistical approach such as the

7

bootstrap. A thorough analysis would involve taking a window of sequence of a certain size, and moving this window along the genome in steps of a defined size, generating perhaps hundreds of trees for visual examination in the process. There are at least two short cuts. One is to analyze only a few windows, defining selected regions according to the output of the diversity analysis. Another is to not examine the entire phylogenetic tree of all subtypes, but to focus on one particular phylogenetic question. Thus, if the initial analyses suggest that a sequence may be a recombinant between two particular subtypes, it is possible to ask simply what is the bootstrap value for the clustering of the new sequence with one or another particular subtype, and plot these values as a function of position along the genome; this is the basis of the "bootscanning" approach (57). Once the subtypes putatively involved in the recombination event have been identified, and the crossover points have been approximately localized, more precisely defined breakpoints can be determined, and their statistical significance assessed, using informative site analysis (19, 52, 53).

Detailed phylogenetic characterization revealed that most of the recombinant viruses have a complex genome structure with multiple points of crossover (7,18,53,60). Some recombinants, like the "subtype E" viruses, which are in fact A/E recombinants (7,18), have a wide-spread geographic dissemination and are responsible for much of the Asian HIV-1 epidemic (69,70). In other areas, recombinants appear to be generated with increasing frequencies as many randomly chosen isolates exhibit evidence of mosaicism (4,8,31,66,71).

Since recombination provides the opportunity for evolutionary leaps with genetic consequences that are far greater than the steady accumulation of individual mutations, the impact of recombination on viral properties must be monitored. Full-length non-recombinant reference sequences for all major HIV-1 groups and subtypes are thus needed to map and characterize the extent of inter-subtype recombination.

Non-subtype B viruses cause the vast majority of new HIV-1 infections worldwide. Although their geographic dissemination is carefully monitored, their immunogenic and biological properties remain largely unknown, in part because well-characterized virological reference reagents are lacking. In particular, full length clones and sequences are rare, since subtype classification is frequently based on small

4

PCR-derived viral fragments. There are currently only five full length, non-recombinant molecular clones available for viruses other than subtype B (45), and these represent only three of the proposed (group M) subtypes (A, C and D). Moreover, only three clones (all derived from subtype D viruses) are replication competent and thus useful for studies requiring functional gene products (45,48,65). Given the unknown impact of genetic variation on correlates of immune protection, subtype specific reagents are critically needed for phylogenetic, immunological and biological studies.

SUMMARY OF INVENTION

The present invention pertains to the isolation and characterization of the genomic sequences of 11 molecular clones for non-subtype B HIV-1 isolates of human immunodeficiency virus type 1 (HIV-1), and nucleic acids derived therefrom. Of these 11 molecular clones, 94IN476.104, 96ZM651.8, and 96ZM751.3 are non-mosaic reference clones of HIV-1 subtype C; 93BR020.1 is a reference clone of HIV-1 subtype F; 90CF056.1 is a reference clone of HIV-1 subtype H; 92RW009.6 is a double recombinant of HIV-1 subtypes A/C; 92NG083.2 and 92NG003.1 are double recombinants of HIV-1 subtypes A/G; 93BR029.4 is a double recombinant of HIV-1 subtypes B/F; 94CY017.41 is a double recombinant of HIV-1 subtype A and a new, as yet undefined, subtype; and 94CY032.3 is a triple recombinant of HIV-1 subtypes A/G/I.

In particular, the present invention relates to nucleic acids comprising the genomic sequences of one or more of these 11 clones for non-subtype B HIV-1 isolates, as well as nucleic acids comprising the complementary (or antisense) sequence of one or more of the genomic sequences of these 11 clones, and nucleic acids derived therefrom.

The invention also relates to vectors comprising the nucleic acid genomic sequence of one or more of these 11 clones, as well as nucleic acids comprising the complementary (or antisense) sequence of one or more of the genomic sequences of these clones, and nucleic acids derived therefrom.

The invention also relates to cultured host cells comprising the nucleic acid genomic sequences of one or more of these 11 clones for non-subtype B HIV-1

isolates, as well as nucleic acids comprising the complementary (or antisense) sequence of one or more of the genomic sequences of these clones, and nucleic acids derived therefrom.

The invention also relates to host cells containing vectors comprising the genomic sequences of one or more of these 11 clones for non-subtype B HIV-1 isolates, as well as nucleic acids comprising the complementary (or antisense) sequence of one or more of the genomic sequences of these clones, and nucleic acids derived therefrom.

The invention also relates to synthetic or recombinant polypeptides encoded by or derived from the nucleic acid sequences of one or more of the genomes of these 11 clones for non-subtype B HIV-1 isolates, and fragments thereof.

The invention also relates to methods for producing the polypeptides of the invention in culture using one or more of these 11 clones for non-subtype B HIV-1 viruses or nucleic acids derived therefrom, including recombinant methods for producing the polypeptides of the invention.

The invention further relates to methods of using the polypeptides of the invention as immunogens to stimulate an immune response in a mammal, such as the production of antibodies, or the generation of cytotoxic or helper T-lymphocytes.

The invention also relates to methods of using the polypeptides of the invention to detect antibodies which immunologically react with non-subtype B HIV-1 viruses in a mammal or in a biological sample.

The invention also relates to kits for the detection of antibodies specific for non-subtype B HIV-1 viruses in a biological sample where said kit contains at least one polypeptide encoded by or derived from the nucleic acid sequences of the invention.

The invention also relates to antibodies, which immunologically react with the virions of one or more of these 11 viruses and/or their encoded polypeptides.

The invention also relates to methods of detecting virions of non-subtype B HIV-1 viruses and/or their encoded polypeptides, or fragments thereof, using antibodies of the invention.

The invention also relates to kits for detecting the virions of non-subtype B HIV-1 viruses and/or their encoded polypeptides, wherein the kit comprises

at least one antibody of the invention.

The invention also relates to a method for detecting the presence of non-subtype B HIV-1 viruses in a mammal or a biological sample, said method comprising analyzing the DNA or RNA of a mammal or a sample for the presence of the RNAs, cDNAs or genomic DNAs which will hybridize to a nucleic acid derived from one or more of these 11 non-subtype B HIV-1 molecular clones. Usually, when a completely complementary probe is used, high stringency conditions are desirable in order to prevent false positives. However, conditions of high stringency should only be used if the probes are complementary to target regions which lack heterogeneity. The stringency of hybridization is determined by a number of factors during hybridization and during the washing procedure, including temperature, ionic strength, length of time, and concentration of formamide, if any. The nucleic acid sequences used in probes should be unique to HIV, i.e., the nucleic acid sequences should be absent from individuals not infected with HIV.

The invention also provides diagnostic kits for the detection of non-subtype B HIV-1 viruses in a mammal using the nucleic acids of the invention. In one embodiment, the kit comprises nucleic acids having sequences useful as hybridization probes in determining the presence or absence of the RNAs, cDNAs or genomic DNAs of non-subtype B HIV-1 viruses. In another embodiment, the kit comprises nucleic acids having sequences useful as primers for reverse-transcription polymerase chain reaction (RT-PCR) analysis of RNA for the presence of HIV-1 viruses in a biological sample.

The invention further relates to isolated and substantially purified nucleic acids, polypeptides and/or antibodies of the invention.

The invention further relates to compositions comprising one or more of the nucleic acids, polypeptides and/or antibodies of the invention.

The invention also relates to computer-generated alignments of the nucleic acid sequences of the viral genomes clones of the 11 clones of this invention, as well as alignments of the encoded amino acid sequences. These sequence alignments serve to highlight regions of homology and non-homology between different sequences and hence, can be used in preparing diagnostic reagents as described herein.

BRIEF DESCRIPTION OF THE FIGURES

INS #1

~~Fig. 1. Phylogenetic relationships of the 11 viral genomes described in~~ this patent application (highlighted) to representatives of all major HIV-1 (group M) subtypes in *gag* (A) and *env* (B) regions. Trees were constructed from full-length *gag* and *env* nucleotide sequences using the neighbor-joining method (see text for details of methodology). Horizontal branch lengths are drawn to scale; vertical separation is for clarity only. Values at the nodes indicate the percent bootstraps in which the cluster to the right was supported (bootstrap values of 75% and higher are shown). Asterisks denote hybrid genomes as determined by additional analyses. Brackets at the right represent the major sequence subtypes of HIV-1 group M. Trees ~~were rooted by using SIVcpzGAB as an outgroup~~.

INS #2

~~Fig. 2. Diversity plots comparing the sequence relationships of the 11~~ viral genomes described in this patent application to each other and to reference sequences from the database. In each of panels A-J, the sequence named above the plots is compared to the sequences listed at the right. U455, LAI, C2220, and NDK are published reference sequences for subtypes A, B, C and D, respectively. Distance values were calculated for a window of 500 bp moved in steps of 10 nucleotides. The x-axis indicates the nucleotide positions along the alignment (gaps were stripped and removed from the alignment). The positions of the start codons of the *gag, pol, vif, vpr, env,* and *nef* genes are shown. The y-axis denotes the distance between the ~~viruses compared (0.05 = 5% divergence).~~

INS #3

~~Fig. 3. Exploratory tree analysis. Neighbor joining trees were~~ constructed for a 500 bp window moved in increments of 100 bp along the multiple genome alignment. Trees depicting discordant branching orders among four of the 11 sequences included in this patent application are shown in panels A-I (hybrid sequences are boxed). The position of each tree in the alignment is indicated; subtypes are identified by brackets. Numbers at nodes indicate the percentage of bootstrap values with which the adjacent cluster is supported (only values above 80% are ~~shown). Branch lengths are drawn to scale.~~

INS #4

~~Fig. 4. Recombination breakpoint analysis for 92RW009.6 and~~ ~~92RW020.4. (A) Bootstrap plots depicting the relationship of 92RW009.6 to~~

H4 Cont

~~representatives of subtype A and C, respectively. Trees were constructed from the~~ multiple genome alignment and the magnitude of the bootstrap value supporting the clustering of 92RW009.6 with U455 and 92UG037.1 (subtype A), or C2220 and 92BR025.8 (subtype C), respectively, was plotted for a window of 500 bp moved in increments of 10 bp along the alignment. Regions of subtype A or C origin are identified by very high bootstrap values (>90%). Points of cross-over of the two curves indicate recombination breakpoints. The beginning of *gag, pol, vif, vpr, env* and *nef* open reading frames are shown. The y-axis indicates the percent bootstrap replicates, which support the clustering of 92RW009.6 with representatives of the respective subtypes. (B) Bootstrap plots depicting the relationship of 93BR029.4 to representatives of subtype Band F, respectively. Analyses are as in (A), except that bootstrap values supporting the clustering of 93BR029.4 with SF2, OYI, MN, LAI and RF (subtype B), or 93BR020.1 (subtype F), respectively, were plotted. Subtype D viruses were excluded from this analysis because of their known close relationship ~~with subtype B viruses.~~

INS H5

~~Fig. 5. Recombination breakpoint analysis of 92NG083.2 and~~ 92NG003.1. Neighbor joining trees depicting discordant branching orders of 92NG003.1 and 92NG083.2 in regions delineated by breakpoints identified by distance plots (not shown) are shown in panels A-D (hybrid sequences are boxed). The position of each tree in the alignment is indicated; subtypes are identified by brackets. Numbers at nodes indicate the percentage of bootstrap values with which the adjacent cluster is supported (only values above 80% are shown). Branch lengths ~~are drawn to scale.~~

Fig. 6. Inferred structure of the five recombinant genomes included in this patent application. LTR sequences were not analyzed and are thus shown as open boxes.

INS H6

~~Fig. 7. Subtype specific genome features. (A) Alignment of deduced~~ Tat (region encoded by second exon) amino acid sequences. Consensus sequences were generated for available representatives of all major subtypes (question marks indicate sites at which fewer than 50% of the viruses contain the same amino acid ~~residue). Dashes denote sequence identity with the consensus sequence, while dots~~

H6 Conv?

~~represent gaps introduced to optimize alignments. A vertical box highlights a~~ premature Tat protein truncation (asterisk) which is present in 11 of 15 subtype D, and 4 of 52 subtype B viruses (frequencies are listed in the column on the right). (B) Alignment of deduced Rev (region encoded by the second exon) protein sequences. ~~(C) Alignment of deduced Vpu protein sequences.~~

Fig. 8: Generation of replication competent proviral clones from long PCR products. The general construction scheme of a replication competent provirus from two separately amplified genomic regions is depicted.

Fig. 9. Diversity plots comparing the sequence relationships of 94CY032.3 to reference sequences from the database. 92UG037.1, LAI, C2220, and ELI are reference sequences for subtypes A, B, C and D, respectively. 92NG083.2 is a known G/A recombinant, but contains only a small subtype A fragment between position 4200 and 4800 (there is presently no full length non-mosaic subtype G reference sequence available). Distance values were calculated for a window of 400 bp moved in steps of 10 nucleotides. The x-axis indicates the nucleotide positions along the alignment (gaps were stripped and removed from the alignment). The positions of the start codons of the *gag, pol, vif, vpr, env,* and *nef* genes are shown. The y-axis denotes the distance between the viruses compared (0.05 = 5% difference).

ens H7

~~Fig. 10. Exploratory tree analysis. Neighbor joining trees were~~ constructed for a 400 bp window moved in increments of 10 bp along the multiple genome alignment. Trees in panel A-K depict the discordant branching orders for 94CY032.3 (highlighted). The position of each tree in the alignment is indicated; subtypes are identified by brackets. Numbers at nodes indicate the percentage of bootstrap values with which the adjacent cluster is supported (only values above 80% ~~are shown). Branch lengths are drawn to scale.~~

Fig. 11. Bootstrap plot analysis to map recombination breakpoints in 94CY032.3. Bootscanning was performed essentially as described, plotting the magnitude of the bootstrap value supporting the clustering of 94CY032.3 with 92UG037.1 (subtype A) in comparison with that of 94CY032.3 and 92NG083.2 ("subtype G") for a window of 400 bp moved in increments of 10 bp along the alignment. Regions of subtype A or G origin are identified by very high bootstrap

values (>80%). The location of eight recombination crossovers is indicated. Breakpoint analysis between position 4200 and 4800 was not possible due to the recombinant nature of 92NG083.2. The beginning of *gag, pol, vif, vpr, env* and *nef* open reading frames are shown. The y-axis indicates the percent bootstrap replicates, which support the clustering of 94CY032.3 with representatives of the respective subtypes.

Fig. 12. Recombination breakpoint analysis of 94CY032.3 in the *vif/vpr* region. Neighbor joining trees depicting the position of 94CY032.3 in regions flanking the breakpoints identified by distance plot analysis (not shown). Trees were constructed from the genomic regions indicated. Subtypes are identified by brackets. Four sequences from Mali represent subtype G (these are the only available subtype G reference sequences in this region, since all other "subtype G" viruses contain A fragments). Numbers at nodes indicate the percentage of bootstrap values with which the adjacent cluster is supported (only values above 80% are shown). Branch lengths are drawn to scale.

INS H8

~~Fig. 13. Nucleotide sequence alignment of the 11 near full-length HIV-1 sequences included in this patent application. Sequences were aligned using CLUSTAL W and adjusted manually using the sequence editor MASE. Dots indicate gaps introduced to optimize the alignment. The beginning and end of all open reading frames are indicated by arrows above or below the alignment. The homologies between the sequences of nucleotides in the eleven independent clones are indicated by dashes. Sequences of nucleotides present uniquely in the various clones (as compared to the corresponding sequences of the other ten clones) are indicated by letters, i.e., the sequences themselves.~~

INS H9

~~Fig. 14. Amino acid sequence alignments of the Gag polypeptides encoded by the 11 near full-length HIV-1 sequences included in this patent application. The homologies between the sequences of amino acids in the various polypeptides encoded by the eleven independent clones are indicated by dashes. Sequences of amino acids present uniquely in the various polypeptides (as compared to the corresponding polypeptides of the other ten clones) are indicated by letters, i.e., the sequences themselves.~~

11

INS H10

~~Fig. 15. Amino acid sequence alignments of the Pol polypeptides~~ encoded by the 11 near full-length HIV-1 sequences included in this patent application. The homologies between the sequences of amino acids in the various polypeptides encoded by the eleven independent clones are indicated by dashes. Sequences of amino acids present uniquely in the various polypeptides (as compared to the corresponding polypeptides of the other ten clones) are indicated by letters, i.e., ~~the sequences themselves~~.

Fig. 16. Amino acid sequence alignments of the Vif polypeptides encoded by the 11 near full-length HIV-1 sequences included in this patent application. The homologies between the sequences of amino acids in the various polypeptides encoded by the eleven independent clones are indicated by dashes. Sequences of amino acids present uniquely in the various polypeptides (as compared to the corresponding polypeptides of the other ten clones) are indicated by letters, i.e., the sequences themselves.

Fig. 17. Amino acid sequence alignments of the Vpr polypeptides encoded by the 11 near full-length HIV-1 sequences included in this patent application. The homologies between the sequences of amino acids in the various polypeptides encoded by the eleven independent clones are indicated by dashes. Sequences of amino acids present uniquely in the various polypeptides (as compared to the corresponding polypeptides of the other ten clones) are indicated by letters, i.e., the sequences themselves.

Fig. 18. Amino acid sequence alignments of the Tat polypeptides encoded by the 11 near full-length HIV-1 sequences included in this patent application. The homologies between the sequences of amino acids in the various polypeptides encoded by the eleven independent clones are indicated by dashes. Sequences of amino acids present uniquely in the various polypeptides (as compared to the corresponding polypeptides of the other ten clones) are indicated by letters, i.e., the sequences themselves.

Fig. 19. Amino acid sequence alignments of the Rev polypeptides encoded by the 11 near full-length HIV-1 sequences included in this patent application. The homologies between the sequences of amino acids in the various

polypeptides encoded by the eleven independent clones are indicated by dashes. Sequences of amino acids present uniquely in the various polypeptides (as compared to the corresponding polypeptides of the other ten clones) are indicated by letters, i.e., the sequences themselves.

Fig. 20. Amino acid sequence alignments of the Vpu polypeptides encoded by the 11 near full-length HIV-1 sequences included in this patent application. The homologies between the sequences of amino acids in the various polypeptides encoded by the eleven independent clones are indicated by dashes. Sequences of amino acids present uniquely in the various polypeptides (as compared to the corresponding polypeptides of the other ten clones) are indicated by letters, i.e., the sequences themselves.

INS H1'

Fig. 21. ~~Amino acid sequence alignments of the Env polypeptides~~ encoded by the 11 near full-length HIV-1 sequences included ~~in this~~ patent application. The homologies between ~~the~~ sequences of amino acids in the various polypeptides encoded ~~by~~ the eleven independent clones are indicated by dashes. Sequences ~~of~~ amino acids present uniquely in the various polypeptides (as compared ~~to~~ the corresponding polypeptides of the other ten clones) are indicated by letters, i.e., ~~the sequences themselves.~~

Fig. 22. Amino acid sequence alignments of the Nef polypeptides encoded by the 11 near full-length HIV-1 sequences included in this patent application. The homologies between the sequences of amino acids in the various polypeptides encoded by the eleven independent clones are indicated by dashes. Sequences of amino acids present uniquely in the various polypeptides (as compared to the corresponding polypeptides of the other ten clones) are indicated by letters, i.e., the sequences themselves.

INS #2 INS F1

DETAILED DESCRIPTION OF THE INVENTION

~~The present invention relates to the determination of the nucleic acid~~ sequences of the complete or near complete genomes ~~of 11~~ non-subtype B HIV-1 viruses isolated from primary ~~isolates~~ collected at major epicenters of the global AIDS pandemic. The ~~nucleotide~~ sequences of these 11 viruses are shown in Fig. 13 (SEQ ID ~~NOS: __ to __)~~

The phrase "derived from" is used throughout the specification and claims with respect to nucleic acids to describe nucleic acid sequences which correspond to a region of the designated nucleotide sequence. Preferably, the sequence of the region from which the nucleic acid is derived is, or is complementary to, a sequence which is unique to the genome of any one of the 11 clones of this invention. However, more preferably, the sequence of the region from which the nucleic acid is derived is, or is complementary to, a sequence which is unique to the viruses in the subtype corresponding to the subtype of any one of the 11 clones of this invention, and whose uniqueness was unknown prior to the disclosure of the clones of this invention. For example, sequences in the Cyprus clone 94CY032.3 which map to the I region are unique wherever they are not identical to known prior art sequences. Whether or not a sequence is unique to the genome of one of the molecular clones or a subtype can be determined by techniques known to those of skill in the art. For example, the sequence can be compared to sequences in databanks, e.g., GenBank, to determine whether it is present in the uninfected host or other organisms. The sequence can also be compared to the known sequences of other viral agents, including other retroviruses. The correspondence or non-correspondence of the derived sequence to other sequences can also be determined by hybridization under the appropriate stringency conditions. Hybridization techniques for determining the complementarity of nucleic acid sequences are well known in the art. In addition, mismatches of duplex polynucleotides formed by hybridization can be determined by known techniques, including for example, digestion with a nuclease such as S1 that specifically digests single-stranded areas in duplex polynucleotides.

Regions of the viral genome from which nucleic acid sequences may be derived include, but are not limited to, regions encoding specific epitopes as well as non-transcribed and non-translated sequences. Preferably, the epitope is unique to HIV viruses in the subtype corresponding to the subtype of the corresponding region of a polypeptide encoded by any one of the 11 clones of this invention, and whose uniqueness was unknown prior to the disclosure of the clones of this invention The uniqueness of the epitope may be determined by its immunological reactivity with HIV viruses of the subtype and lack of immunological reactivity with other HIV viruses of the other subtypes. Methods for determining immunological reactivity are

known in the art, e.g., radioimmunoassay and ELISA and other assays mentioned herein. The uniqueness of an epitope can also be determined by computer searches of known databases, e.g., for the polynucleotide sequences which encode the eptiope, and by amino acid sequence comparisons with other known proteins.

The derived nucleic acid is not necessarily physically derived from the nucleotide sequence shown, but may be generated in any manner, including for example, chemical synthesis or DNA replication or reverse transcription or transcription, which are based on the information provided by the sequence of bases in the region(s) from which the nucleic acid is derived. The derived nucleic acid is comprised of at least 6-12 bases, more preferably at least 15-19 bases, more preferably at least 30 bases. The derived nucleic acid may also be larger, e.g., at least 100 bases in length, depending on the desired use of the nucleic acid. In addition, regions or combinations of regions corresponding to that of the designated sequence may be modified in ways known in the art to be consistent with an intended use. The derived nucleic acid may be a polynucleotide or polynucleotide analog.

The term "recombinant nucleotide" or "recombinant nucleic acid" as used herein intends a nucleic acid of genomic, cDNA, semisynthetic, or synthetic origin which, by virtue of its origin or manipulation: (1) is not associated with all or a portion of the nucleic acid with which it is associated in nature; and/or (2) is linked to a nucleic acid other than that to which it is linked in nature.

The term "polynucleotide" as used herein refers to a polymeric form of nucleotides of any length, either ribonucleotides or deoxyribonucleotides. This term refers only to the primary structure of the molecule. Thus, this term includes double- and single-stranded DNA, as well as double- and single-stranded RNA. It also includes modified, for example, by methylation and/or by capping, and unmodified forms of the polynucleotide.

INS F2

INS A13

~~The present invention relates to nucleic acids having the genomic sequence of any one of the 11 molecular clones for non-subtype HIV-1 isolates of this invention as shown in Fig. 13 (SEQ ID NOS:___ to ___), as well as fragments (or partial sequences) thereof. The invention also relates to nucleic acids having complementary (or antisense) sequences to the sequences shown in Fig. 13 (SEQ ID NOS:___ to ___), as well as fragments (or partial sequences) thereof. Partial~~

427568_1

15

H13 CONT.

~~sequences may be obtained by various methods, including restriction digestion of nucleic acids with sequences shown in Fig. 13 (SEQ ID NOS:___ to ___), PCR amplification, and direct synthesis. Partial sequences may be all or part of the LTR and/or other untranslated regions of the genomes of one or more of the 11 viral clones of this invention, and/or all or part of the genes encoding the Gag, Pol, Vif, Vpr, Env, Tat, Rev, Nef and Vpu proteins and/or complementary (or antisense) sequences thereof. Nucleic acids of the invention also include cDNA, mRNA, and other nucleic acids derived from the genomic sequences of one or more of these 11 HIV-1 clones. Sequences of the genes encoding Gag, Pol, Vif, Vpr, Env, Tat, Rev, Nef and Vpu are identified in Fig. 13.~~

Genomic sequences of seven of the 11 clones of the invention have been made publicly available. The GenBank Accession numbers are as follows:

Clone	Accession No.	Sequence ID No.
92RW009.6	U88823	
92NG003.1	U88825	
92NG083.2	U88826	
93BR020.1	AF005494	
93BR029.4	AF005495	
90CF056.1	AF005496	
94CY032.3	AF049337	
94CY017.41	-	
96ZM651.8	-	
96ZM751.3	-	
94IN476.104	-	

The nucleic acids of the invention may be present in vectors or host cells in tissue culture or other media. The nucleic acids of the invention may also be isolated and substantially purified by methods known in the art.

Nucleic acids of about 17 bases to about 35 bases in length are particularly preferred for use as primers in PCR amplification (see, *e.g.*, the primers UP1A and R/U5 (17mer and 22mer, respectively) and UP1AMlu1 and Low1Mlu1 (28mer and 35mer respectively)). Nucleic acids of about 14 to about 25 bases in

427568_1

16

length are particularly preferred for use in nucleotide arrays. (See, eg., ref. 108, which uses 20 to 25 mers).

The present invention also relates to vectors and host cells comprising the nucleic acids of the invention.

The present invention also relates to compositions comprising one or more of the nucleic acids, vectors, and/or host cells of the invention.

The present invention further relates to methods of using the nucleic acids, vectors, and/or host cells of the invention, and/or compositions thereof. For example, the invention relates to the use of nucleic acids of the invention as diagnostic agents to detect the presence or absence of non-subtype B HIV-1 viruses in a sample.

The present invention also relates to a method for detecting the presence of HIV-1 viruses which are related to the viruses of this invention in a mammal, using the nucleic acids of this invention.

In one embodiment, the detection method involves analyzing DNA obtained from a mammal suspected of harboring HIV-1 viruses. DNA can be isolated by methods well known in the art.

The methods for analyzing the DNA for the presence of the viruses of this invention include Southern blotting (86), dot and slot hybridization (87), and nucleotide arrays (*see, e.g.*, US 5,445,934 and US 5,733,729).

(INS F3

H14

~~The nucleic acid probes used in the detection methods set forth above~~ are derived from nucleic acid ~~sequences~~ shown in Fig. 13 (SEQ ID NOS:___ to ___). The size of such probes ~~is~~ at least 10-12 bases long, more usually at least about 19 bases long, ~~more~~ usually from about 200 to about 500 bases, and often exceeding ~~about 1000 bases.~~

The nucleic acid probes of this invention may be DNA or RNA. Nucleic acids can be synthesized using any of the known methods of nucleotide synthesis (see, e.g., refs. 88, 89, 90), or they can be isolated fragments of naturally occurring or cloned DNA. In addition, those skilled in the art would be aware that nucleotides can be synthesized by automated instruments sold by a variety of manufacturers or can be commercially custom ordered and prepared. The probes of this invention may also be nucleotide analogs, such as nucleotides linked by phosphodiester, phosphorothiodiester, methylphosphonodiester, or

methylphosphonothiodiester moieties (91) and peptide nucleic acids (PNAs), in which the sugar-phosphate backbone of the polynucleotide is replaced with a polyamide or "pseudopeptide" backbone (92).

The nucleic acid probes can be labeled using methods known to one skilled in the art. Such labeling techniques can include radioactive labels, biotin, avidin, enzymes and fluorescent molecules (93).

INS #15

~~The nucleic acid probes used in the detection methods set forth above are derived from~~ sequences substantially homologous to one or more of the sequences shown in Fig. 13 (SEQ ID NOS:___ to ___), or their complementary sequences. By "substantially homologous", as used throughout the specification and claims to describe the nucleic acid sequence of the present invention, is meant a high level of homology between the nucleic acid sequence and one or more of the sequences of Fig. 13 (SEQ ID NOS:___ to ___), or its complementary sequence. Preferably, the level of homology is in excess of 80%, more preferably in excess of 90%, with a preferred nucleic acid sequence being in excess of 95% homologous with a portion of one or more of the sequences shown in Fig. 13 (SEQ ID NOS:___ to ___), or its complement. The size of such probes is usually at least 20 nucleotides, more usually from about 200 ~~to 500 nucleotides, and often exceeding 1000 nucleotides.~~

Although complete complementarity is not necessary, it is preferred that the probes are made completely complementary to the corresponding portion of the genome, mRNA or cDNA target of at least one of the 11 viruses of this invention.

The probes can be packaged into diagnostic kits. Diagnostic kits may include ingredients for labeling and other reagents and materials needed for the particular hybridization protocol in addition to the probes.

In another embodiment of the invention, the detection method comprises analyzing the RNA of a mammal for the presence of HIV-1 viruses which are related to one or more of the 11 the viruses of this invention. RNA can be isolated by methods well known in the art.

INS #16

~~The methods for analyzing the RNA for the presence of the viruses of this invention include Northern blotting (94), dot and slot hybridization, filter hybridization (95), RNase protection (93), and reverse-transcription polymerase chain reaction (RT-PCR) (96). A preferred method is RT-PCR. In this method, the RNA~~

#16 Cont

~~can be reverse transcribed to first strand cDNA using a nucleic acid primer or primers~~ derived from one or more of the nucleotide sequences shown in Fig. 13 (SEQ ID NOS:___ to ___). Once the cDNAs are synthesized, PCR amplification is carried out using pairs of primers designed to hybridize with sequences in the genomes of one or more of the non-subtype B HIV-1 viruses of this invention which are an appropriate distance apart (at least about 50 bases) to permit amplification of the cDNA and subsequent detection of the amplification product. Each primer of a pair is a single-stranded nucleic acid of about 20 to about 60 bases in length where one primer (the "upstream" primer) is complementary to the original RNA and the second primer (the "downstream" primer) is complementary to the first strand of cDNA generated by reverse transcriptions of the RNA. The target sequence is generally about 100 to about 300 bases in length but can be as large as 500-1500 bases or more, e.g., 9,000 bases. Optimization of the amplification reaction to obtain sufficiently specific hybridization to the nucleotide sequences of these viruses is well within the skill in the art and is ~~preferably achieved by adjusting the annealing temperature.~~

The amplification products of PCR can be detected either directly or indirectly. In one embodiment, direct detection of the amplification products is carried out via labeling of primer pairs. Labels suitable for labeling the primers of the present invention are known to one skilled in the art and include radioactive labels, biotin, avidin, enzymes and fluorescent molecules. The desired labels can be incorporated into the primers prior to performing the amplification reaction. Alternatively, the desired labels can be incorporated into the primer extension products during the amplification reaction in the form of one or more labeled dNTPs. In one embodiment of the present invention, the labeled amplified PCR products can be detected by agarose gel electrophoresis followed by ethidium bromide staining and visualization under ultraviolet light or via direct sequencing of the PCR-products. The labeled amplified PCR products can also be detected by binding to immobilized oligonucleotide arrays.

In yet another embodiment, unlabelled amplification products can be detected via hybridization with labeled nucleic acid probes in methods known to one skilled in the art, such as dot or slot blot hybridization or filter hybridization.

The invention also relates to methods of using these nucleic acids to

produce polypeptides *in vitro* or *in vivo*.

In one embodiment of the invention, a recombinant method of making a polypeptide of the invention comprises:

(a) preparing a nucleic acid capable of directing a host cell to produce a polypeptide encoded by the genome of any one of the non-subtype B HIV-1 viruses of this invention;

(b) cloning the nucleic acid into a vector capable of being transferred into and replicated in a host cell, such vector containing operational elements for expressing the nucleic acid, if necessary;

(c) transferring the vector containing the nucleic acid and operational elements into a host cell capable of expressing the polypeptide;

(d) growing the host under conditions appropriate for expression of the polypeptide; and

(e) harvesting the polypeptide.

The present invention also relates to non-recombinant methods of making the polypeptides and nucleic acids of the invention. In addition to synthetic methods, the non-recombinant methods involve culturing the viruses of this invention in cell lines, preferably in uninfected human peripheral blood mononuclear cells, under conditions appropriate for expression of the polypeptides and nucleic acids. This invention thus also relates to the polypeptides and nucleic acids produced by the virus in cell culture. The polypeptides and nucleic acids may be isolated and purified by methods known in the art.

The vectors contemplated for use in the present invention include any vectors into which a nucleic acid sequence as described above can be inserted, along with any preferred or required operational elements, and which vector can then be subsequently transferred into a host cell and, preferably, replicated in such cell. Preferred vectors are those whose restriction sites have been well documented and which contain the operational elements preferred or required for transcription of the nucleic acid sequence. Vectors may also be used to prepare large amounts of nucleic acids of the invention, which may be used, e.g., to prepare probes or other nucleic acid constructs.

When expression of a polypeptide is desired, the "operational

elements" as discussed herein include at least one promoter sequence capable of initiating transcription of the nucleic acid sequence, at least one leader sequence, at least one terminator codon and/or termination signal, and any other DNA sequences necessary or preferred for appropriate transcription and subsequent translation of the vector nucleic acid. In particular, it is contemplated that such vectors will preferably contain at least one origin of replication recognized by the host cell along with at least one selectable marker.

Preferred expression vectors of this invention are those which function in bacterial and/or eukaryotic cells. Examples of vectors which function in eukaryotic cells include, but are not limited to Venezuelan equine encephalitis virus vectors, simian virus vectors, vaccinia virus vectors, adenovirus vectors, herpes virus vectors, or vectors based on retroviruses, such as murine leukemia virus, or HIV or other lentivirus (97).

The selected expression vector may be transfected into a suitable bacterial or eukaryotic cell system for purposes of expressing the recombinant polypeptide. Eukaryotic cell systems include but are not limited to cell lines such as HeLa, COS-1, 293T , MRC-5, or CV-1 cells. Primary human cells, such as lymph node cells, macrophages, etc., are also useful in practicing the invention.

The expressed polypeptides may be detected directly by methods known in the art including, but not limited to, Coomassie blue staining and Western blotting or indirectly, such as in detection of the expression product of a reporter gene, such as luciferase.

In another embodiment of the invention, the method comprises administering a composition comprising a vector comprising a nucleic acid of the invention to a mammal to produce a polypeptide *in vivo*.

The present invention also relates to polypeptides encoded by and/or derived from the nucleotide sequences of this invention. These polypeptides may be natural, synthetic or produced by recombinant methods. Polypeptides can be obtained as a crude lysate or can be purified by standard protein purification procedures known in the art which may include differential precipitation, molecular size exclusion chromatography, ion-exchange chromatography, isolectric focusing, gel electrophoresis and affinity and immunoaffinity chromatography. The polypeptides

may be purified by passage through a column containing a resin which has bound thereto antibodies specific for an open reading frame (ORF) polypeptide. The present invention also relates to compositions comprising one or more of the polypeptides of the invention.

A polypeptide or amino acid sequence derived from a designated nucleic acid sequence refers to a polypeptide having an amino acid sequence identical to that of a polypeptide encoded by the sequence, or a portion thereof wherein the portion consists of at least 6-8 amino acids, and more preferably at least 10 amino acids, and more preferably at least 11-15 amino acids, and most preferably at least 30 amino acids or which is immunologically cross-reactive with a polypeptide encoded by the sequence. The polypeptide may also be larger, e.g., at least 100 amino acids in length, depending on the desired use of the polypeptide. Polypeptides from the V3-loop region and the "crown" of gp41 of Env are particularly preferred.

A recombinant or derived polypeptide is not necessarily translated from a designated nucleic acid sequence; it may be generated in any manner, including for example, chemical synthesis, or expression of a recombinant expression system, or isolation from any of the 11 HIV-1 viruses of this invention.

It should be noted that the nucleotide sequences described herein represent one embodiment of the present invention. Due to the degeneracy of the genetic code, it is to be understood that numerous choices of nucleotides may be made that will lead to a sequence capable of directing production of the polypeptides set forth above. As such, nucleic acid sequences which are functionally equivalent to the sequences described herein are intended to be encompassed within the present invention. For example, preferred codons which are appropriate to the host cell may be used (*see, e.g.*, WO 98/34640), or the sequence may be modified to reduce the effect of any inhibitory/instability sequences and to provide for Rev-independent gene expression. (98).

~~The polypeptides of this invention consist of at least 6-12 amino acids~~, more preferably at least 13-18 amino acids, even more preferably ~~at least~~ 19-24 amino acids and most preferably ~~at least 25-30 amino~~ acids encoded by, or otherwise derived ~~from, any one of the genomic sequences shown in Fig. 13 (SEQ ID NOS: __ to __).~~

The present invention further relates to the use of polypeptides of the

invention as diagnostic agents.

In one embodiment, the polypeptides of the invention can be used in immunoassays for detecting the presence of antibodies against non-subtype B HIV-1 viruses in a mammal and for diagnosing the presence of infection of any of these viruses in a mammal.

For the purposes of the present invention, "mammal" as used throughout the specification and claims, includes, but is not limited to humans, chimpanzees, mangabeys, other other primates.

In a preferred embodiment, test serum is reacted with a solid phase reagent having a surface-bound polypeptide of this invention as an antigen. The solid surface reagent can be prepared by known techniques for attaching polypeptides to solid support material. These attachment methods include non-specific adsorption of the polypeptide to the support or covalent attachment of the polypeptide to a reactive group on the support. After reaction of the antigen with an antibody against any one of the viruses of this invention in the serum, unbound serum components are removed by washing and the antigen-antibody complex is reacted with a secondary antibody such as labeled anti-human antibody. The label may be an enzyme which is detected by incubating the solid support in the presence of a suitable fluorimetric or colorimetric reagent. Other detectable labels may also be used, such as radiolabels or colloidal gold, and the like.

Immunoassays of the present invention may be a radioimmunoassay, Western blot assay, immunofluorescent assay, enzyme immunoassay, chemiluminescent assay, immunohistochemical assay and the like. Standard techniques for ELISA are well known in the art. Such assays may be a direct, indirect, competitive, or noncompetitive immunoassay as described in the art (*see, e.g.,* ref. 99). Biological samples appropriate for such detection assays include, but are not limited to serum, liver, saliva, lymphocytes or other mononuclear cells.

Polypeptides of the invention may be prepared in the form of a kit, alone, or in combinations with other reagents such as secondary antibodies, for use in immunoassays.

In yet another embodiment, the polypeptides of the invention can be used as immunogens to raise antibodies and/or stimulate cellular immunity in a

mammal.

The immunogen may be a partially or substantially purified peptide. Alternatively, the immunogen may be a cell, cell lysate from cells transfected with a recombinant expression vector, or a culture supernatant containing the expressed polypeptide. The immunogen may comprise one or more structural proteins, and/or one or more non-structural proteins of the HIV-1 clones of this invention, or a mixture thereof.

The effective amount of polypeptide per unit dose sufficient to induce an immune response depends, among other things, on the species of mammal inoculated, the body weight of the mammal and the chosen inoculation regimen, as well as the presence or absence of an adjuvant, as is well known in the art. Inocula typically contain polypeptide concentrations of about 1 microgram to about 50 milligrams per inoculation (dose), preferably about 10 micrograms to about 10 milligrams per dose, most preferably about 100 micrograms to about 5 milligrams per dose.

The term "unit dose" as it pertains to the inocula refers to physically discrete units suitable as unitary dosages for mammals, each unit containing a predetermined quantity of active material (polypeptide) calculated to produce the desired immunogenic effect in association with the required diluent.

Inocula are typically prepared as a solution in a physiologically acceptable carrier such as saline, phosphate-buffered saline and the like to form an aqueous pharmaceutical composition.

The route of inoculation of the polypeptides of the invention is typically parenteral and is preferably intramuscular, sub-cutaneous and the like. The dose is administered at least once. In order to increase the antibody level, at least one booster dose may be administered after the initial injection, preferably at about 4 to 6 weeks after the first dose. Subsequent doses may be administered as indicated.

To monitor the antibody response of individuals administered the compositions of the invention, antibody titers may be determined. In most instances it will be sufficient to assess the antibody titer in serum or plasma obtained from such an individual. Decisions as to whether to administer booster inoculations or to change the amount of the composition administered to the individual may be at least partially

based on the titer.

The titer may be based on an immunobinding assay which measures the concentration of antibodies in the serum which bind to a specific antigen. The ability to neutralize *in vitro* and *in vivo* biological effects of the viruses of this invention may also be assessed to determine the effectiveness of the immunization.

For all therapeutic, prophylactic and diagnostic uses, the polypeptide of the invention, alone or linked to a carrier, as well as antibodies and other necessary reagents and appropriate devices and accessories may be provided in kit form so as to be readily available and easily used.

Where immunoassays are involved, such kits may contain a solid support, such as a membrane (e.g., nitrocellulose), a bead, sphere, test tube, microtiter well, rod, and so forth, to which a receptor such as an antibody specific for the target molecule will bind. Such kits can also include a second receptor, such as a labeled antibody. Such kits can be used for sandwich assays. Kits for competitive assays are also envisioned.

The immunogens of this invention can also be generated by the direct administration of nucleic acids of this invention to a subject. DNA-based vaccination has been shown to stimulate humoral and cellular responses to HIV-1 antigens in mice (100-103) and macaques (103, 104). More recent studies in infected chimpanzees have shown a possible application of this strategy in HIV-1-infected humans: DNA vaccination of HIV-1-infected chimpanzees with a construct that drives expression of HIV-1 *env* and *rev* appeared well-tolerated, and immunized animals demonstrated a boost in antibody response followed by a >1 log decrease in their virus loads (104). A DNA-based vaccine containing HIV-1 *env* and *rev* genes was injected into HIV-infected human patients in three doses (30, 100 or 300 micrograms) at 10-week intervals. Increased antibodies against gp120 were observed in the 100 and 300 μg groups. Increases were also noted in cytotoxic T lymphocyte (CTL) activity against gp160-bearing targets and in lymphocyte proliferative activity (105, 106). DNA-based vaccines containing HIV *gag* genes, with modification of the viral nucleotide sequence to incorporate host-preferred codons (*see, e.g.*, WO 98/34640), and/or to reduce the effect of inhibitory/instability sequences (*see, e.g.*, ref. 98), have likewise been described.

Therefore, it is anticipated that the direct injection of RNA or DNA vectors of this invention encoding viral antigen can be used for endogenous expression of the antigen to generate the viral antigen for presentation to the immune system without the need for self-replicating agents or adjuvants, resulting in the generation of antigen-specific CTLs and protection from a subsequent challenge with a homologous or heterologous strain of virus.

CTLs in both mice and humans are capable of recognizing epitopes derived from conserved internal viral proteins and are thought to be important in the immune response against viruses. By recognition of epitopes from conserved viral proteins, CTLs may provide cross-strain protection. CTLs specific for conserved viral antigens can respond to different strains of virus, in contrast to antibodies, which are generally strain-specific.

Thus, direct injection of RNA or DNA encoding the viral antigen has the advantage of being without some of the limitations of direct peptide delivery or viral vectors (*see, e.g.*, ref. 107 and the discussions and references therein). Furthermore, the generation of high-titer antibodies to expressed proteins after injection of DNA indicates that this may be a facile and effective means of making antibody-based vaccines targeted towards conserved or non-conserved antigens, either separately or in combination with CTL vaccines targeted towards conserved antigens. These may also be used with traditional peptide vaccines, for the generation of combination vaccines. Furthermore, because protein expression is maintained after DNA injection, the persistence of B and T cell memory may be enhanced, thereby engendering long-lived humoral and cell-mediated immunity.

Nucleic acids encoding a polypeptide of this invention can be introduced into animals or humans in a physiologically or pharmaceutically acceptable carrier using one of several techniques such as injection of DNA directly into human tissues; electroporation or transfection of the DNA into primary human cells in culture (*ex vivo*), selection of cells for desired properties and reintroduction of such cells into the body, (said selection can be for the successful homologous recombination of the incoming DNA to an appropriate preselected genomic region); generation of infectious particles containing the *gag* and/or other genes encoded by the viruses of this invention, infection of cells *ex vivo* and reintroduction of such cells into the body; or

direct infection by said particles *in vivo*. Substantial levels of polypeptide will be produced leading to an efficient stimulation of the immune system.

Also envisioned are therapies based upon vectors, such as viral vectors containing nucleic acid sequences coding for the polypeptides described herein. These molecules, developed so that they do not provoke a pathological effect, will stimulate the immune system to respond to the polypeptides.

The effective amount of nucleic acid immunogen per unit dose to induce an immune response depends, among other things, on the species of mammal inoculated, the body weight of the mammal and the chosen inoculation regimen, as is well known in the art. Inocula typically contain nucleic acid concentrations of about 1 microgram to about 50 milligrams per inoculation (dose), preferably about 10 micrograms to about 10 milligrams per dose, most preferably about 100 micrograms to about 5 milligrams per dose.

Immunization can be conducted by conventional methods. For example, the immunogen can be used in a suitable diluent such as saline or water, or complete or incomplete adjuvants. Further, the immunogen may or may not be bound to a carrier. While it is possible for the immunogen to be administered in a pure or substantially pure form, it is preferable to present it as a pharmaceutical composition, formulation or preparation.

The formulations of the present invention, both for veterinary and for human use, comprise an immunogen as described above, together with one or more physiologically or pharmaceutically acceptable carriers and optionally other therapeutic ingredients. The carrier(s) must be "acceptable" in the sense of being compatible with the other ingredients of the formulation and not deleterious to the recipient thereof. The formulations may conveniently be presented in unit dosage form and may be prepared by any method well-known in the pharmaceutical art. The immunogen can be administered by any route appropriate for antibody production such as intravenous, intraperitoneal, intramuscular, subcutaneous, and the like. The immunogen may be administered once or at periodic intervals until a significant titer of antibody against any of the 11 viruses of this invention is produced. The antibody may be detected in the serum using an immunoassay. The host serum or plasma may be collected following an appropriate time interval to provide a composition

comprising antibodies reactive with the virus particle or encoded polypeptide. The gamma globulin fraction or the IgG antibodies can be obtained, for example, by use of saturated ammonium sulfate or DEAE Sephadex, or other techniques known to those skilled in the art.

In addition to its use to raise antibodies, the administration of the immunogens of the present invention may be for use as a vaccine for either a prophylactic or therapeutic purpose. When provided prophylactically, a vaccine(s) of the invention is provided in advance of any exposure to any one or more of the 11 non-subtype B viruses of this invention or in advance of any symptoms due to infection of these viruses. The prophylactic administration of a vaccine(s) of the invention serves to prevent or attenuate any subsequent infection of these viruses in a mammal. When provided therapeutically, a vaccine(s) of the invention is provided at (or shortly after) the onset of infection or at the onset of any symptom of infection or any disease or deleterious effects caused by these viruses. The therapeutic administration of the vaccine(s) serves to attenuate the infection or disease. The vaccine(s) of the present invention may, thus, be provided either prior to the anticipated exposure to the viruses of this invention or after the initiation of infection.

In another embodiment, the polypeptides of the invention can be used to prepare antibodies against epitopes of the viruses of this invention that are useful in diagnosis.

The term "antibodies" is used herein to refer to immunoglobulin molecules and immunologically active portions of immunoglobulin molecules. Exemplary antibody molecules are intact immunoglobulin molecules, substantially intact immunoglobulin molecules and portions of an immunoglobulin molecule, including those portions known in the art as Fab, Fab', $F(ab')_2$ and F(v) as well as chimeric antibody molecules.

An antibody of the present invention is typically produced by immunizing a mammal with an immunogen or vaccine of the invention. In one embodiment, the immunogen or vaccine contains one or more polypeptides of the invention, or a structurally and/or antigenically related molecule, to induce, in the mammal, antibody molecules having immunospecificity for the immunizing peptide or peptides. The peptide(s) or related molecule(s) may be monomeric, polymeric,

conjugated to a carrier, and/or administered in the presence of an adjuvant. In another embodiment, the immunogen or vaccine contains one or more nucleic acids encoding one or more polypeptides of the invention, or one or more nucleic acids encoding structurally and/or antigenically related molecules, to induce, in the mammal, the production of the immunizing peptide or peptides. The antibody molecules may then be collected from the mammal if they are to be used in immunoassays or for providing passive immunity.

The antibody molecules of the present invention may be polyclonal or monoclonal. Monoclonal antibodies may be produced by methods known in the art. Portions of immunoglobulin molecules may also be produced by methods known in the art.

The antibody of the present invention may be contained in various carriers or media, including blood, plasma, serum (e.g., fractionated or unfractionated serum), hybridoma supernatants and the like. Alternatively, the antibody of the present invention is isolated to the extent desired by well known techniques such as, for example, by using DEAE SEPHADEX, or affinity chromatography. The antibodies may be purified so as to obtain specific classes or subclasses of antibody such as IgM, IgG, IgA, IgG_1, IgG_2, IgG_3, IgG_4 and the like. Antibody of the IgG class are preferred for purposes of passive protection.

The presence of the antibodies of the present invention, either polyclonal or monoclonal, can be determined by, but are not limited to, the various immunoassays described above.

The antibodies of the present invention have a number of diagnostic and therapeutic uses. The antibodies can be used as an *in vitro* diagnostic agent to test for the presence of any one or more of the 11 HIV-1 viruses of this invention in biological samples in standard immunoassay protocols. Preferably, the assays which use the antibodies to detect the presence of these viruses in a sample involve contacting the sample with at least one of the antibodies under conditions which will allow the formation of an immunological complex between the antibody and the viral antigen that may be present in the sample. The formation of an immunological complex if any, indicating the presence of one or more of these viruses in the sample, is then detected and measured by suitable means. Such assays include, but are not

limited to, radioimmunoassays, (RIA), ELISA, indirect immunofluorescence assay, Western blot and the like. The antibodies may be labeled or unlabeled depending on the type of assay used. Labels which may be coupled to the antibodies include those known in the art and include, but are not limited to, enzymes, radionucleotides, fluorogenic and chromogenic substrates, cofactors, biotin/avidin, colloidal gold and magnetic particles. Modification of the antibodies allows for coupling by any known means to carrier proteins or peptides or to known supports, for example, polystyrene or polyvinyl microtiter plates, glass tubes or glass beads and chromatographic supports, such as paper, cellulose and cellulose derivatives, and silica.

Such assays may be, for example, of direct format (where the labeled first antibody reacts with the antigen), an indirect format (where a labeled second antibody reacts with the first antibody), a competitive format (such as the addition of a labeled antigen), or a sandwich format (where both labeled and unlabelled antibody are utilized), as well as other formats described in the art. In one such assay, the biological sample is contacted with antibodies of the present invention and a labeled second antibody is used to detect the presence of any one of the HIV-1 viruses of this invention, to which the antibodies are bound.

The antibodies of the present invention are also useful as a means of enhancing the immune response.

The antibodies may be administered with a physiologically or pharmaceutically acceptable carrier or vehicle therefor. A physiologically acceptable carrier is one that does not cause an adverse physical reaction upon administration and one in which the antibodies are sufficiently soluble and retain their activity to deliver a therapeutically effective amount of the compound. The therapeutically effective amount and method of administration of the antibodies may vary based on the individual patient, the indication being treated and other criteria evident to one of ordinary skill in the art. A therapeutically effective amount of the antibodies is one sufficient to reduce the level of infection by one or more of the viruses of this invention or attenuate any dysfunction caused by viral infection without causing significant side effects such as non-specific T cell lysis or organ damage.

The route(s) of administration useful in a particular application are apparent to one or ordinary skill in the art. Routes of administration of the antibodies

include, but are not limited to, parenteral, and direct injection into an affected site. Parenteral routes of administration include but are not limited to intravenous, intramuscular, intraperitoneal and subcutaneous.

The present invention includes compositions of the antibodies described above, suitable for parenteral administration including, but not limited to, pharmaceutically acceptable sterile isotonic solutions. Such solutions include, but are not limited to, saline and phosphate buffered saline for intravenous, intramuscular, intraperitoneal, or subcutaneous injection, or direct injection into a joint or other area.

Antibodies for use to elicit passive immunity in humans are preferably obtained from other humans previously inoculated with pharmaceutical compositions comprising one or more of the polypeptides of the invention. Alternatively, antibodies derived from other species may also be used. Such antibodies used in therapeutics suffer from several drawbacks such as a limited half-life and propensity to elicit an immune response. Several methods are available to overcome these drawbacks. Antibodies made by these methods are encompassed by the present invention and are included herein. One such method is the "humanizing" of non-human antibodies by cloning the gene segment encoding the antigen binding region of the antibody to the human gene segments encoding the remainder of the antibody. Only the binding region of the antibody is thus recognized as foreign and is much less likely to cause an immune response.

In providing the antibodies of the present invention to a recipient mammal, preferably a human, the dosage of administered antibodies will vary depending upon such factors as the mammal's age, weight, height, sex, general medical condition, previous medical history and the like.

In general, it is desirable to provide the recipient with a dosage of antibodies which is in the range of from about 5 mg/kg to about 20 mg/kg body weight of the mammal, although a lower or higher dose may be administered. In general, the antibodies will be administered intravenously (IV) or intramuscularly (IM).

The invention also relates to the use of antisense nucleic acids to inhibit translation of peptides encoded by the HIV-1 viruses of this invention. The antisense nucleic acids are complementary to the viral mRNAs encoding peptides of this invention. The antisense nucleic acids may be in the form of synthetic nucleic acids or

they may be encoded by a nucleotide construct, or they may be semi-synthetic. The antisense nucleic acids may be delivered to the cells using methods known to those skilled in the art.

Kits designed for diagnosis of the HIV-1 viruses of this invention in a biological sample can be constructed by packaging the appropriate materials, including the nucleic acids and/or polypeptides of this invention and/or antibodies which specifically react with antigens of one or more of these viruses, along with other reagents and materials required for the particular assay.

INS F7

INS H18

~~The present invention further relates to computer-generated alignments~~ of any one or more of the nucleotide sequences shown ~~in Fig. 13~~ (SEQ ID NOS:___ to ___). Computer analysis ~~of the nucleotide~~ sequences, such as the one shown in Fig. 13, can ~~be carried~~ out using commercially available computer program known to one ~~skill in the art.~~

INS F6

INS H19

~~In one embodiment, the sequences shown in Fig. 13 (SEQ ID NOS:~~___ to ___) are aligned by the computer program CLUSTAL ~~(67)~~ and adjusted with multiple-aligned sequence editor (12). ~~The~~ computer analysis results in the distribution of 11 ~~sequences~~ into various genotypes. Five of these sequences represent non-~~recombinant~~ members of HIV-1 subtypes, and the other six sequences represent ~~HIV-1 intersubtype recombinants.~~

The grouping of the molecular clones into mosaic and non-mosaic genotypes is shown below:

Name of Clone	Genotypes
94CY017.41	A/?
94IN476.104	C
96ZM651.8	C
96ZM751.3	C
93BR020.1	F
90CF056.1	H
92RW009.6	A/C
92NG083.2	A/G
92NG003.1	A/G
93BR029.4	B/F
94CY032.3	A/G/I

For those sequences representing recombinant members of HIV-1, a variety of phylogenetic methods were used to further characterize the subtype composition.

~~The multiple computer-generated alignments of nucleotide sequences~~ are shown in Figure 13. The multiple computer-generated alignments of encoded amino acid sequences are shown in Figures 14-22. These alignments serve to highlight regions of homology and non-homology between different sequences and hence, can be used by one skilled in the art to design oligonucleotides and ~~polypeptides useful as reagents in diagnostic assays for HIV-1~~.

The following examples illustrate certain embodiments of the present invention, but should not be construed as limiting its scope in any way. Certain modifications and variations will be apparent to those skilled in the art from the teachings of the forgoing disclosure and the following examples, and these are intended to be encompassed by the spirit and scope of the invention.

A330

#7

20

EXAMPLE 1

Materials and Methods

Virus isolates

All viruses used were propagated in normal donor peripheral blood mononuclear cells (PBMCs) and thus represent primary isolates. Their biological phenotype (SI/NSI), year of isolation, relevant epidemiological and clinical information, as well as appropriate references are summarized in Table 1. For consistency, isolates are labeled according to WHO nomenclature (28). Preliminary subtype classification was made on the basis of partial *env* and/or *gag* gene sequences (1,17,19,43).

Amplification of near complete HIV-1 genomes using long PCR methods

INS F9

(Near) full length HIV-1 genomes were amplified from short-term cultured PBMC DNA essentially as described (18,56) using the GeneAmp XL kit (Perkin Elmer Cetus, Foster City, Calif.) and primers spanning the tRNA primer binding site (upstream primer UP1A: 5'-AGTGGCGCCCGAACAGG-3') (SEQ ID NO: ___) and the R/U5 junction in the 3' long terminal repeat (downstream primer Low2: 5'-TGAGGCTTAAGCAGTGGGTTTC-3') (SEQ ID NO: ___). Some isolates were amplified with primers containing Mlu1 restriction enzyme sites to facilitate subsequent subcloning into plasmid vectors (upstream primer UP1AMlu1: 5'-TCTCTacgcgtGGCGCCCGAACAGGAC-3' (SEQ ID NO: ___); downstream primer Low1Mlu1: 5'- ACCAGacgcgtACAACAGACGGGCACACACTA-CTT-3' (SEQ ID NO: ___); lower case letters indicate the Mlu1 restriction site). Whenever possible, PBMC DNAs were diluted prior to PCR analysis to attempt amplification from single proviral templates. Cycling conditions included a hot start (94°C, 2 min), followed by 20 cycles of denaturation (94°C; 30 sec) and extension (68°C; 10 min), followed by 17 cycles of denaturation (94°C; 30sec) and extension (68°C, 10min) with 15 second increments per cycle. PCR products were visualized by agarose gel electrophoresis and subcloned into pCRII by T/A overhang or following cleavage with Mlu1 into a modified pTZ18 vector (pTZ18Mlu1) containing a unique Mlu1 site in its polylinker. Transformations were performed in INVαF' cells, and colonies were screened by restriction enzyme digestion for full length inserts (transformation

efficiencies were generally poor, yielding only a few recombinant colonies; however, once subcloned, full length genomes were stable in their respective vectors). One full length clone per isolate was randomly chosen for subsequent sequence analysis.

Construction of a full length and infectious molecular clone of 94UG114.1

A 674 bp fragment spanning most of the viral LTR (lacking 1-92 of U3 sequences) as well as the untranslated leader sequence preceding *gag*, was amplified from 94UG114 PBMC DNA, using primers and conditions described previously (18). After sequence confirmation, this LTR fragment was cloned into the pTZ18Mlu1 vector, which was subsequently cleaved with *Nar*1 (in the primer binding site) and *Mlu*1 (in the polylinker) to allow the insertion of the 94UG114.1 long PCR product cleaved with the same restriction enzymes. The resulting plasmid clone comprised a full length 94UG114.1 genome with 3' and 5' LTR fragments containing all regulatory elements necessary for viral replication. A similar strategy could be used to construct replication competent genomes for all 11 clones reported in this application.

Sequence analysis of HIV-1 genomes

A number of the clones described herein were sequenced using the shotgun sequencing approach (37). Briefly, viral genomes were released from their respective plasmid vectors by cleavage with the appropriate restriction enzymes, purified by gel electrophoresis, and sonicated (Model XL2020 Sonicator; Heat System Inc., Framingdale, N.Y.) to generate randomly sheared DNA fragments 600-1,000 bp in length. Following purification by gel electrophoresis, fragments were end-repaired using T4 DNA polymerase and Klenow enzyme and ligated into SmaI digested and dephosphorylated M13 or pTZ18 vectors. Approximately 200 shotgun clones were sequenced for each viral genome using cycle sequencing and dye terminator methodologies on an automated DNA sequenator (Model 377A; Applied Biosystems, Inc.). Sequences were determined for both strands of DNA. Other clones were sequenced directly using the primer walking approach (primers were designed approximately every 300 bp along the genome for both strands). Proviral contigs were assembled from individual sequences using the SEQUENCHER program (Gene Codes Corporation, Ann Arbor, Mich.). Sequences were analyzed using EUGENE (Baylor College of Medicine, Houston, TX) and MASE (12).

Phylogenetic tree analysis

Phylogenetic relationships of the newly derived viruses were estimated from sequence comparisons with previously reported representatives of HIV-1 group M (45). Multiple *gag* and *env* sequence alignments were obtained from the Los Alamos sequence database (http://hiv-web.lanl.gov/HTML/alignments.html). Newly derived *gag* and *env* sequences were added to these alignments using the CLUSTAL W profile alignment option (67) and adjusted manually using the alignment editor MASE (12). All partial sequences were removed from these alignments. Sites where there was a gap in any of the remaining sequences, as well as areas of uncertain alignment, were excluded from all sequence comparisons. Pairwise evolutionary distances were estimated using Kimura's two parameter method to correct for superimposed substitutions (26). Phylogenetic trees were constructed using the neighbor-joining method (55), and the reliability of topologies was estimated by performing bootstrap analysis using 1,000 replicates (13). NJPLOT was used to draw trees for illustrations (49). Phylogenetic relationships were also determined using maximum-parsimony (with repeated randomized input orders; ten iterations) as well as maximum-likelihood approaches, implemented using the programs DNAPARS and DNAML from the PHYLIP package (14).

Complete genome alignment

All newly derived HIV-1 genome sequences were aligned with previously reported (45) full length representatives of HIV-1 subtype A (U445), B (LAI, RF, OYI, MN, SF2), C (C2220), D (ELI, NDK, Z2Z6), and "E" (90CF402.1, 93TH253.3, CM240) as well as SIVcpzGAB as an outgroup using the CLUSTAL W (67) profile alignment option (the alignment includes the untranslated leader sequence, *gag, pol, vif, vpr, tat, rev, vpu, env, nef* and available 3' LTR sequences). Sequences that needed to be excluded from any particular analysis were removed only after gap-tossing was performed on the complete alignment containing all sequences. This ensured that all positions were comparable in different runs with different sequences.

Diversity plots

The percent diversity between selected pairs of sequences was determined by moving a window of 500 bp in 10 bp increments along the genome

alignment. The divergence values for each pairwise comparison were plotted at the midpoint of the 500 bp segment.

Bootstrap plots

Bootscanning was performed on neighbor-joining trees using SEQBOOT, DNADIST (using Kimura's correction), NEIGHBOR and CONSENSUS from the PHYLIP package (14) for a window of 500 bp moved along the alignment in increments of 10 bp. 1000 replicates were evaluated for each phylogeny. The program ANALYZE from the bootscanning package (57) was used to examine the clustering of the putative hybrid with representatives of the subtypes presumed to have been involved in the recombination event. The bootstrap values for these sequence were plotted at the midpoint of each window.

Exploratory tree analysis

Exploratory tree analysis was performed using the bootstrap plot approach described above, except in this case an increment of 100 bp was used and each neighbor-joining tree was viewed using DRAWTREE from the PHYLIP package (14). In addition, all full length sequences (except known recombinants) were included into the analysis.

Informative site analysis

To estimate the location and significance of cross-overs, each putative hybrid sequence was compared with a representative of each of the two subtypes inferred to have been involved in the recombination event, and an appropriate outgroup. Recombination breakpoints were mapped by examining the linear distribution of phylogenetically informative sites supporting the clustering of the hybrid with each of the two "parental" subtypes, essentially as described (52,53). Potential breakpoints were inserted between each pair of adjacent informative sites, and the extent of heterogeneity between the two sides of the breakpoint, with respect to numbers of the two kinds of informative site, was calculated as a 2 x 2 chi square value; the likely breakpoint was identified as that which gave the maximal chi-square value. Since the alignments contained more than one putative cross-over, this analysis was performed looking for one and two breakpoints at a time, and repeated on

subsections of the alignment defined by breakpoints already identified. To assess the probability of obtaining (by chance) chi-square values as high as those observed, 10,000 random permutations of the informative sites were examined

Nucleotide sequence accession numbers

GenBank accession numbers for several of the (near) full length HIV-1 proviral sequences disclosed in this application are listed in Table 2, and are hereby incorporated by reference.

EXAMPLE 2

Identification of non-subtype B HIV-1 viruses

Molecular cloning of non-subtype B HIV-1 isolates

Of the geographically diverse HIV-1 isolates described herein, five had previously been classified as members of (group M) subtypes A (92RW009), F (92BR020, 92BR029), and G (92NG003, 92NG083) on the basis of *env* (17,19) and/or *gag* sequences (1). One (90CF056) was chosen because it originated from a major epicenter of the African AIDS epidemic. In addition, 90CF056 was of interest because it did not fall into any known subtype at the time of its first genetic characterization (43). Isolates from Zambia (96ZM651 and 96ZM751) and India (94IN476) were chosen because of the known subtype C prevalence in those countries. The two isolates from Cyprus (94CY017 and 94CY032) were selected because of the extensive diversity of HIV-1 in the drug user population (29). Table 1 summarizes available demographic and clinical information, as well as biological data concerning the isolate phenotype (SI/NSI). Only viruses grown in normal donor PBMCs were selected for analysis.

T0390

Table 1. Epidemiological and clinical information for study isolates

Isolate[a]	Sex[b]	Age	City	Country	Risk factor[c]	Disease status[d]	Antiviral therapy	Year of isolation	Source[e]	Biological phenotype[f]	Preliminary subtype assignment	Refs.
94CY017.41	F	35	Nicosia	Cyprus	Het	SM	n/a	1994	ADARC	n/a	A/?	29
94CY032.3	M	35	Nicosia	Cyprus	Het	AS	n/a	1994	ADARC	n/a	G/A/I	29
96ZM651.8	M	47	Lusaka	Zambia	Het	SM	n/a	1996	UAB	n/a	n/a	n/a
96ZM751.3	M	26	Lusaka	Zambia	Het	SM	No	1996	UAB	n/a	n/a	n/a
94IN476.104	F	n/a	Pune	India	n/a	n/a	No	1994	ADARC	n/a	n/a	n/a
93BR020	M	52	Rio de Janeiro	Brazil	Bi	AS	No	1993	WHO	SI	F	19, 72
90CF056 (U4056)	M	n/a	Bangui	CAR	Het	AS	No	1990	PIB	NSI	U	43
92RW009	F	24	Kigali	Rwanda	Het	AS	No	1992	WHO	NSI	A[h]	17, 72
93BR029	M	17	Sao Paulo	Brazil	n/a	AS	No	1993	WHO	NSI	F[h]	19, 72
92NG083 (JV1083)	F	27	Jos	Nigeria	n/a	AIDS	No	1992	IHV	NSI	G[h]	1
92NG003 (G3)	F	24	Jos	Nigeria	Het	AS	n/a	1992	IHV	NSI	G[h]	1

[a] Isolates were named according to WHO nomenclature (previous designations are listed in parentheses).

[b] M, male; F, female.

[c] Het, heterosexual contact; Bi, bisexual contact; Hemo, hemophiliac patient.

[d] AS, asymptomatic; SM, symptomatic.

[e] TJU, Thomas Jefferson University, Philadelphia, PA; PIB, Pasteur Institute, Bangui, CAR; IHV, Institute of Human Virology, Baltimore, MD; WHO, World Health Organization, Geneva, Switzerland; UAB, University of Alabama.

[f] Determined in MT-2 assay as described (72); NSI, non-syncytium inducing; SI, synctium inducing.

[g] n/a, information not available.

[h] Isolates identified to be recombinant in present study.

427568_1

39

The viral genomes were cloned by long PCR methods using primers homologous to the tRNA primer binding site (upstream primer) and the polyadenylation signal in the 3' LTR (downstream primer). This amplification strategy generated (near) full length genomes containing all coding and regulatory regions, except for 70 to 80 bps of 5' unique LTR sequences (U5). All isolates, regardless of subtype classification, yielded long PCR products with the same set of primer pairs. In some instances, genomes were amplified with primers containing *Mlu*1 restriction enzyme sites. This greatly facilitated subsequent subcloning into a plasmid vector (Table 2).

Sequence analysis of (near) full length HIV-1 genomes

All eleven HIV-1 genomes were sequenced in their entirety using either shotgun sequencing or primer walking approaches. The long PCR derived clones ranged in size from 8,952 to 8,999 base pairs, and spanned the genome from the primer binding site to the R/U5 junction of the 3' LTR. Inspection of potential coding regions revealed that all clones contained the expected reading frames for *gag*, *pol*, *vif*, *vpr*, *tat*, *rev*, *vpu*, *env* and *nef*. In addition, all major regulatory sequences, including promotor and enhancer elements in the LTR, the packaging signal, splice sites, etc., appeared to be intact. None of the genomes had major deletions or rearrangements, although inspection of the deduced protein sequences identified inactivating mutations in seven of the eleven clones (Table 2). However, most of these were limited to point mutations in single genes and were thus amenable to repair. Only two genomes (92NG003.1 and 92NG083.2) contained stop codons, small deletions and frameshift mutations in several genes, rendering them multiply defective. Importantly, no inactivating mutations were identified in 93BR020.1 (subtype F), 90CF056.1 (subtype H), and 96ZM651.8 (subtype C), suggesting that these clones encoded biologically active genomes (Table 2). Nucleic acids containing repaired coding sequences, as well as the polypeptides encoded by the repaired coding sequences, are also considered to be a part of the invention.

T0410

Table 2. Inactivating mutations in near-complete HIV-1 genomes

Clone	Defective gene(s)	Inframe stop codon[a]	Frameshift mutation[a]	Altered initiation codon[a]	Plasmid vector[d]	GenBank accession number
93BR020.1	none	-	-	-	pCR2.1	AF005494
90CF056.1	none	-	-	-	pCR2.1	AF005496
92RW009.6	*gag*	-	213	-	pTZ18 (Mlu1)	U88823
93BR029.4	*gag*	-	260,472	-	pTZ18 (Mlu1)	AF005495
92NG083.2	*gag, vpu*	360	5462[b]	157	pTZ18 (Mlu1)	U88826
92NG003.1[c]	*vpr, vpu, nef*	-	5024[b], 5485[b]	8113	pTZ18 (Mlu1)	U88825
96ZM651.8	none	-	-	-	pTZ18 (Mlu1)	pending
96ZM751.3	*gag/pol/env*	7567	1067/2688	-	pTZ18 (Mlu1)	pending
94IN476.104	*pol/vpr*	3021	-	-	pTZ18 (Mlu1)	pending
94CY032.3	*vif/env/vpr*	4518/7125	5199	-	pTZ18 (Mlu1)	pending
94CY017.41	*rev*	-	-	5327	pCRII	pending

[a] Numbers indicate the position of the inactivating mutation within the sequence.
[b] Frameshift mutations associated with more extensive nucleotide sequence deletions (10-16 bp).
[c] 92NG003.1 also has a 33 bp deletion in the V3 loop region of *env*.
[d] Genomes were either subcloned by T/A overhang into pCRII, or via Mlu1 sites in the primer sequences into pTZ18 (Mlu1).

41

INS
H21

EXAMPLE 3

Phylogenetic analyses in *gag* and *env* regions

~~To determine the phylogenetic relationships of the viruses described~~ herein, evolutionary trees from full length *gag* and *env* sequences were first constructed. This was done to confirm the authenticity of previously characterized strains, classify the new viruses, and compare viral branching orders in trees from two genomic regions. The results confirmed a broad subtype representation among the selected viruses (Fig. 1). Strains fell into six of the seven major (non-B) clades, including three for which full length sequences are not available (i.e., F, G and H). However, comparison of the *gag* and *env* topologies also identified three strains with discordant branching orders. 92RW009.6 grouped with subtype C viruses in *gag*, but with subtype A viruses in *env*. Similarly, 93BR029.4 clustered with subtype B viruses in *gag*, but with subtype F viruses in *env*. 94CY017.41 appeared to cluster within subtype A viruses in *env*, but fell into an unknown subtype in *gag*. However, characterization of the latter strain is still ongoing. These different phylogenetic positions were supported by high bootstrap values and thus indicated that these strains ~~were intersubtype recombinants.~~

EXAMPLE 4

Diversity plots

To characterize the putative recombinants as well as the other strains in regions outside *gag* and *env*, pairwise sequence comparisons with available full length sequences from the database were performed. A multiple genome alignment was generated which included the new sequences as well as U455 (subtype A), LAI, RF, OYI, MN and SF2 (subtype B), C2220 (subtype C), ELI, NDK and Z2Z6 (subtype D), and 90CF402.1, 93TH253.3 and CM240 ("subtype E"). The percent nucleotide sequence diversity between sequence pairs was then calculated for a window of 500 bp moved in steps of 10 bp along the alignment. Importantly, distance values were calculated only after all sites with a gap in any of the sequences were removed from the alignment. This ensured that all comparisons were made across the same sites.

Fig. 2 depicts selected distance plots for the newly characterized viruses. For example in Fig. 2A, 93BR020.1 (putative subtype F) is compared to U455 (subtype A), NDK (subtype D), C2220 (subtype C) and 90CF056.1 (putative subtype H). The resulting plots all exhibit very similar diversity profiles characterized by alternating regions of sequence variability and conservation (values range from 7% divergence near the 5' and 3' ends of *pol*, to 30% in the segment of *env* encoding the V3 region). Moreover, the four plots are virtually superimposable, indicating that 92BR020.1 is roughly equidistant from U455, NDK, C2220 and 90CF056.1 over the entire length of its genome. A very similar set of distance curves was also obtained from comparisons of 94CY017.41 with 90CF056.1, 92BR025.8, 93BR020.1, U455, and NDK (Fig. 2B), and from comparisons of both 93BR020.1 and 90CF056.1 with representatives of subtype B and "E" (data not shown). These results indicating that 93BR020.1 and 90CF056.1 are equidistant from each other as well as from members of subtypes A, B, C, D and "E", together with the *gag* and *env* phylogenetic trees (Fig. 1), suggest that 93BR020.1 and 90CF056.1 represent non-recombinant members of subtypes F and H, respectively.

Very similar data were also obtained when 90CF056.1 was subjected to diversity plot analysis using the same set of reference sequences (Fig. 2F). Again, distance curves exhibited very similar profiles indicating approximate equidistance among the strains analyzed, except when viruses from the same subtype were compared. For example, in Fig. 2C distances between 94IN476.104 (putative subtype C) and U455, 93BR020.1, 90CF056.1, NDK and 92BR025.8, respectively, are depicted. As expected, the 92BR025.8 (putative subtype C) plot falls clearly below all others, indicating the lower level of sequence divergence between viruses from the same subtype (ranging from about 4% in *pol* to about 17% in *env*). Importantly, however, inter- and intra- diversity plots follow each other very closely, i.e., the same genomic regions exhibit proportionally higher and lower levels of divergence. See also the diversity plot analysis for 92ZM651.8 (Fig. 2G) and 96ZM751.3 (Fig. 2H). Thus, both at the level of inter- and intra-subtype comparisons, there was no evidence of mosaicism in the genomes of these three viruses. Together with the results in Fig. 1, this suggests that strains 94IN476.104, 96ZM651.8 and 96ZM751.3 represent non-mosaic members of subtype C.

By contrast, the diversity plots of the putative recombinants 92RW009.6 (Fig. 2D) and 93BR029.4 (Fig. 2I) exhibited disproportionate levels of sequence divergence from different subtypes along their genome, consistent with their discordant branching orders in *gag* and *env* trees. As shown in Fig. 2D, 92RW009.6 is most similar to the subtype C strain C2220 in the 5' half of *gag*, most of *pol*, *vif*, *vpr*, as well as *nef* (the C2220 curve falls below all others). However, in the 3' end of *gag*, the 5' end of *pol*, and most of *env*, 92RW009.6 is most similar to the subtype A strain U455 (the U455 curve falls below all others). Similarly in Fig. 2I, 93BR029.4 is most similar to the subtype B strain LAI in *gag*, *pol* and *vpr*, while it is most similar to the putative subtype F strain 93BR020.1 in *vif*, *env* and *nef* regions. In each case, the magnitude of the difference between the new sequence and the most similar subtype was no greater than the diversity seen within subtypes. Thus, these data suggest that 92RW009.6 and 93BR029.1 represent mosaics, comprised of subtypes A/C and B/F, respectively. In each case, the plots suggested several (at least four) cross-overs; these are the minimum number of recombination breakpoints, since the window size used makes it unlikely that recombinant regions shorter than 500 bp would be detected.

Finally, inspection of the diversity plots for 92NG003.1 (Fig. 2J) and 92NG083.2 (Fig. 2E) also revealed disproportionate levels of sequence variation, although not as pronounced as for 92RW009.6 and 93BR029.4. Isolates 92NG003.1 and 92NG083.2 are equidistant from members of subtypes A-F and H for the most part of their genome, suggesting that they represent an independent subtype, i.e., subtype G. However, in the *vif/vpr* region the U455 distance plot falls below all others, suggesting a disproportionately closer relationship to subtype A. Assuming that U455 is non-mosaic, these results suggest that both 92NG003.1 and 92NG083.2 contain short fragments of subtype A sequence in the central region of their genome.

EXAMPLE 5

Exploratory tree analyses

INS H23

~~To examine the phylogenetic position of the newly derived strains~~ relative to each other and to the reference sequences ~~over the~~ entire genome, exploratory ~~tree~~ analyses were performed using the same multiple genome alignment ~~generated for the diversity plots (Fig. 3). A total of 79 trees were constructed for~~

44

overlapping fragments of 500 bp, moved in 100 bp increments along the alignment. As expected, four genomes were identified that clustered in different subtypes in different parts of their genome. These included 93BR029.4 which alternated between subtypes F and B, 92RW009.6 which alternated between subtypes A and C, and 92NG083.2 and 92NG003.1 which grouped either independently or within subtype A. Interestingly, the latter two strains exhibited distinct patterns of mosaicism. In trees spanning the region 3501-4000, 92NG003.1 clustered within subtype A, while 92NG083.2 clustered independently, presumably representing subtype G. In contrast to these strains, there was no evidence for a hybrid genome structure in 94IN476.104, 96ZM651.8, 96ZM751.3, 93BR020.1 or 90CF056.1. These viruses branched consistently in all regions analyzed. Based on these findings and the results from the diversity plots, it appeared that five of the eleven selected HIV-1 strains represent non-recombinant reference strains for subtypes C (94IN476.104, 96ZM651.8, 96ZM751.3), F (93BR020.1) and H (90CF056.1), respectively, while at least five are intersubtype recombinants. CY017.41 may be recombinant, but work is in progress in this regard.

A23 Cont

EXAMPLE 6

Recombination breakpoint analysis

To map the location of the recombination breakpoints in 92RW009.6 and 93BR029.4, bootstrap plots and informative site analyses were used (18,52,53). Unrooted trees were constructed which included U455, 92UG037.1, LAI, MN, OYI, SF2, RF, C2220, 92BR025.1, NDK, ELI, Z2Z6, 93BR020.1 and 90CF056.1; then the magnitude of the bootstrap values supporting (i) the clustering of 92RW009.6 with members of subtype A (U455, 92UG037.1) or C (2220, 92BR025.8), as well as (ii) the clustering of 93BR029.4 with members of subtype B (LAI, MN, OYI, MN, RF) or F (92BR020.1) was determined (in the latter case subtype D viruses were excluded because of their known close relationship to subtype B viruses). Fig. 4 depicts the results of 797 such phylogenetic analyses generated for each genome, performed on a window of 500 nucleotides moved in steps of 10 nucleotides. Very high bootstrap values (> 80%) supporting the clustering of 92RW009.6 with subtype C were apparent in *gag*, the 3' two thirds of *pol*, and *nef*. By contrast, significant branching of

ins 24

45

#24
Conv7

~~92RW009.6 with subtype A was apparent in the *gag/pol* overlap and the *env* region.~~ In a small region (4,000 to 4,200) in the middle of the genome, 92RW009.6 appeared not to cluster significantly with either subtype, but further inspection revealed that this was due to a small number of informative sites. These data thus indicated four points of recombination crossovers between subtypes A and C (Fig. 4A). A similar analysis identified six recombination breakpoints between subtypes B and F in 93BR029.4 (Fig. 4B). These included two more (in *gag*) than were apparent from the diversity ~~plot analysis (compare Fig. 2), indicating a greater sensitivity of this approach~~.

To map the recombination cross-over points in 92RW009.6 and 93BR029.1 more precisely, the distribution of phylogenetically informative sites supporting alternative tree topologies were examined (52,53). Briefly, this was done in a four sequence alignment which included the query sequence, a representative of each of the two subtypes presumed to have been involved in the recombination event, and an outgroup. Breakpoints were identified by looking for statistically significant differences in the ratios of sites supporting one topology versus another. Consistent with the bootscanning data, this analysis identified four breakpoints in 92RW009.6, and six in 93BR029.4 (Table 3). A schematic representation of the mosaic genomes of 92RW009.6 and 93BR029.4 is depicted in Figure 6.

09520171.081498

46

Table 3. Informative site analysis of 92RW009.6 and 93BR029.4

Clone	Region#	Subtype	Informative Sites		
			subtype A (U455)	subtype C (C2220)	outgroup (NDK)
92RW009.6	1-1037	C	8	32	8
	1085-1940	A	17	5	4
	1986-5288	C	18	99	27
	5293-7238	A	60	9	13
	7254-8431	C	12	55	12
			subtype B (LAI)	subtype F (93BR020)	outgroup (C2220)
93BR029.4	1-735	B	18	6	3
	755-896	F	1	10	0
	930-4247	B	99	10	14
	4340-4668	F	2	15	1
	4787-5166	B	15	0	5
	5244-8242	F	15	139	13
	8250-8429	B	13	0	0

Numbers mark positions in the four sequence alignment which includes the untranslated leader sequence (1-120), *gag* (121-1537), *pol* (1370-4340), *vif* (4285-4856), *vpr* (4799-5073), the first *tat* exon (5054-5271), *vpu* (5276-5488), *env* (5406-7726), *nef* (7727-8313) and the 3′ LTR (7991-8468). Note that position 8468 does not correspond to the end of the LTR but is the last position in the alignment after gaps have been tossed. The 5′ LTR is not included in the alignment.

ins H25

~~Because of the lack of a full length subtype G reference sequence,~~ recombination breakpoint analysis of 92NG003.1 and 92NG083.2 required a different approach. The analyses summarized in Fig. 2 and Fig. 3 suggested that these two viruses contained subtype A sequences in the middle of their genome. To attempt to confirm this, and to define the extent of these putative subtype A fragments, a more detailed diversity plot analysis of the viral middle region (between position 3,000 and 6,000) was performed using different viral strains and varying window sizes (ranging from 200 to 400 bp) to examine the extent of sequence divergence of 92NG083.2 and 92NG003.1 from members of other subtypes, including subtype A. Diversity plots for 92NG003.1 compared to U455, C2220, NDK and 92NG083.2 and for 92NG083.2 ~~compared to U455, C2220, NDK and 92NG003.1 depicted representative results.~~

T0470

47

H25 CONT

~~(using a window size of 300 bp moved in steps of 10 bp along the alignment) (data not~~ shown). Similar to the data shown in Fig. 2, the two "subtype G" viruses are roughly equidistantly related to members of subtypes A (U455), C (C2220), and D (NDK), except for two regions in 92NG003.1 and one region in 92NG083.2 where both viruses are disproportionately more closely related to U455 than they are to each other. Noting the points at which the "G"-A distance increases or decreases relative to the others allowed the tentative identification of recombination breakpoints. For example, at position 3400, the U455 plot falls whereas the C2220, NDK and 92NG083.2 plots do not, and around site 3600 the U455 plot crosses the 92NG083.2 plot. Bearing in mind the window size of 300 nucleotides, this finding suggested that a recombination cross-over occurred around position 3500. Similar "G"-A plot crossings around positions 3800, 4200 and 5200 (in the diversity plot for 92NG003.1), and around positions 4200 and 4800 (in the diversity plot for 92NG083.2), suggested additional ~~recombination breakpoint~~s.

INS H26

~~Phylogenetic trees were then constructed using the regions of sequence~~ defined by these putative breakpoints (Fig. 5). This analysis generally supported the conclusions drawn from the diversity plots, i.e., 92NG003.1 clustered with subtype A viruses in the region between 3501 and 3800, whereas 92NG083.2 did not; and both 92NG003.1 and 92NG083.2 clustered with subtype A viruses in the region 4201 and 4800. However, neither the diversity plot nor the tree analysis allowed the definition of the boundaries of the subtype A fragments with certainty. Nevertheless, the data indicated that (i) both 92NG083.2 and 92NG003.1 represent G/A recombinants, (ii) that they are the result of different recombination events because some of their breakpoints are clearly different, and (iii) that 92NG083.2 likely encodes a non-recombinant *pol* gene. A schematic representation of the mosaic genomes of ~~92NG083.2 and 92NG003.1 is shown in Fig. 6.~~

EXAMPLE 7

Subtype specific genome features

INS H27

~~Having classified the new viruses with respect to their subtype~~ assignments, their sequences ~~were examined~~ for clade-specific signature sequences. ~~Comparing deduced amino acid sequences gene by gene, several subtype speci~~fic

48

A27 Cont

~~features were found (Fig. 7). For example, most subtype D viruses contain an in-~~ frame stop codon in the second exon of *tat,* which removes 13 to 16 amino acids from the carboxy terminus of the Tat protein (Fig. 7A). Similarly, all subtype C viruses (including 94IN476.104, 96ZM651.8 and 96ZM751.3) contain a stop codon in the second exon of *rev* which would be predicted to shorten this protein by 16 amino acids (Fig. 7B). Subtype C viruses also contain a 15 base pair insertion at the 5' end of the *vpu* gene (Fig. 7C) which extends the putative membrane spanning domain of the Vpu protein by 5 amino acids (data not shown). Although these changes are unlikely to alter the function of the respective gene products in a major way (e.g., the known functional domains of both Tat and Rev proteins are not affected by these changes), it is possible that they could influence their mechanism of action in a subtle (but ~~nevertheless biologically important) manner.~~

Of the eleven non-subtype B clones identified herein, phylogenetic analysis identifies five of these viruses as non-recombinant members of subtypes C (three), F and H, which increases the number of non-subtype B reference strains available. Among these, the (near) full length genomes of 93BR020.1 and 90CF056.1 represent the first such strains for subtypes F and H, respectively. Five of the other viruses were found to represent complex mosaics of subtypes A and C, A and G (two), B and F and A, G and I. One, 94CY017.41, is not yet fully characterized. Both A/G recombinants originated from Nigeria, but must have arisen from independent recombination events since they are not closely related and differ in their patterns of mosaicism. One of these (92NG083.2) appears to contain only a single short (perhaps 600bp) segment of subtype A origin in the *vif/vpr* region, and in the absence of (as yet) any full length subtype G virus, thus serves as a (non-mosaic) subtype G representative for *gag, pol, env*, and *nef* regions. Importantly, the genomes were generated in such a way that they can be tested for biological activity following a simple reconstruction step. An example of such a reconstructed genome giving rise to replication competent virus (94UG114.1) demonstrates that this approach is feasible. See “Materials and Methods,” supra, and the schematic diagram in Figure 8.

Given the apparent prevalence of mosaic viruses, it is clear that subtype specific reference strains can only be defined as such after comprehensive recombination analysis. Small subgenomic fragments or even full length *gag* and *env*

sequences are not sufficient to identify all hybrid genomes. Although multiple cross-overs are a characteristic feature of retroviral recombination and have been found in many of the mosaic HIV-1 genomes examined (7,19,53,60,62), the examples of 92NG003.1 and 92NG083.2 demonstrate that cross-overs may be confined to regions outside of *gag* and *env*. Thus, elimination of the possibility that a virus is recombinant requires the determination of substantial (if not all) portions of its genome. As a consequence, subtype specific reference reagents, such as immunogens for cross-clade CTL and neutralization assays, should be derived from viral isolates for which a complete genome has been characterized.

These considerations emphasize the need for detailed analyses using reliable methods for identification of recombinant viral sequences. The above results indicate that diversity plots, depicting the distance between the query sequence and a set of reference sequences in moving windows along the genome, represent an excellent initial screening tool. The extent of sequence divergence (between any pair of viruses) varies along the genome, but since all plots are shown in the same graph, particular regions where the query sequence is anomalously highly similar to (or divergent from) other sequences can be readily identified. For example, this approach uncovered the subtype A-like regions in the middle of the putative "subtype G" genomes 92NG003.1 and 92NG083.2 (Figs 2J and 2E; Fig. 5). However, the results from such analyses relying only on extents of sequence divergence must be treated with some caution, because they are susceptible to variation in evolutionary rate in different lineages. Once suspicious regions have been identified, phylogenetic analyses of windows of sequence around these regions can be used to look for discordant branching orders, and to identify the subtypes likely to have been involved in the recombination event. The bootstrap value supporting the clustering of the query sequence with sequences of the supposed "parental" subtypes can be examined, again in moving windows along the genome. Finally, informative site analysis can be used to map as precisely as possible the breakpoints of the putative recombination events (52,53).

Clearly, recombination analysis relies on the availability of accurately defined non-mosaic reference sequences. Thus, location of the breakpoints in the two G/A recombinant viruses identified here must remain tentative because of the lack of

such reference sequences for subtype G. The precise positions of breakpoints in the recently characterized Thai and CAR "subtype E" viruses are similarly uncertain (7,18), in this case for lack of a complete non-mosaic subtype E reference sequence. It should also be emphasized that currently designated reference sequences may require revision in the future. For example, the inadvertent inclusion of recombinant "reference" sequences in previous tree analyses (19,40) led to an incorrect subtype assignment of subtype G and "E" gp41 sequences. As more sequences become available, it is thus possible that one or more of the viral sequences currently designated as non-recombinant may be identified as a hybrid.

Example 8

Identification of the HIV-1 Clone 94CY032.3

Full length reference clones and sequences are currently available for eight HIV-1 group M subtypes (A - H), but none have been reported for subtypes I and J, which have only been identified in a handful of individuals. Phylogenetic information for subtype I, in particular, is limited since only a very small *env* gene fragment (400 bp in the C2-V3 region) obtained from only two individuals (a heterosexual couple of intravenous drug users from Cyprus) has been analyzed. To characterize subtype I in greater detail, long range PCR was employed to clone a full length provirus (94CY032.3) from a short-term cultured isolate (94CY032) established from one of the two individuals originally reported to be infected with this subtype.

7WB F10

Using primers homologous to the tRNA primer binding site (5'-TCTCT-acgcgtGGCGCCCGAACAGGGAC-3' (SEQ ID NO: ____), lower case letters indicate an Mlu1 site) and the polyadenylation signal in the 3' LTR (5'-ACCAGacgcgtACAACAGACGGG-CACACACTACTT-3') (SEQ ID NO: ____), long range PCR was used to amplify near full length genomic fragments, which contained all coding and regulatory regions except for 102 bp of 5' unique LTR sequences (U5) (for methodological details concerning the long range PCR approach see refs. 18, 56, 79). Amplification products were subcloned into an a plasmid vector, mapped by restriction enzyme digestion, and one clone (94CY032.3) was

51

selected for further analysis. A 694 bp fragment spanning the remainder of the LTR was amplified separately using a semi-nested approach (18).

The complete sequence of 94CY032.3 was determined using the primer walking approach [GenBank accession numbers: AF049337 (genome) and AF049338 (LTR)]. Examination of potential coding regions revealed the expected reading frames for *gag*, *pol*, *vif*, *vpr*, *tat*, *rev*, *vpu*, *env* and *nef* (Fig. 13). None of the genes contained major deletions, insertions or rearrangements. However, both *env* and *vif* genes contained single in-frame stop codons (Fig. 13). There was also a frameshift at position 5199 (single base pair insertion) which altered the C-terminus (last six amino acid residues) of the Vpr protein. All other protein domains of known function as well as major regulatory sequences, including the primer binding site, the packaging signal and major splice sites, appeared to be intact. Similarly, the number, position and consensus sequences of promoter and enhancer elements in the 94CY032.3 LTR were indistinguishable from those of most other HIV-1 strains, except for the presence of an unusual TATA sequence (TAAAA), thus far only found in "subtype E" (A/E) viruses from Thailand and the Central African Republic (7, 18).

To compare 94CY032.3 to previously reported subtype I sequences, a phylogenetic tree was constructed from C2-V3 sequences, including representatives of all 10 known group M subtypes (data not shown). As expected, 94CY032.3 clustered most closely with CYHO321 and CYHO322, sequences amplified from *uncultured* PBMC DNA of the same individual (HO32) from whom the 94CY032 isolate was derived. 94CY032.3 also clustered very closely with CYHO311, a sequence derived from the sexual partner of HO32 (29), strongly suggesting that the two infections were epidemiologically linked. Finally, as observed in the past (29), all subtype I sequences clustered independently, forming a distinct lineage roughly equidistant from all other subtypes, including subtype J (30). These findings thus confirmed the authenticity of the 94CY032.3 clone and validated it as a representative of subtype I in the C2-V3 region of the viral envelope.

To characterize the remainder of the 94CY032.3 genome, pairwise sequence comparisons were then performed with recently reported non-mosaic reference sequences for subtypes A-H (32, 79) as well as selected intersubtype recombinants (83). This approach has been useful for identifying regions of unusual

sequence similarity (or dissimilarity) as an indicator of recombination (18, 79). Briefly, 94CY032.3 was added (using the profile alignment option of CLUSTAL W; 27) to a multiple genome alignment which included a total of 28 sequences from the database (81) representing subtypes A (U455, 92UG037.1), B (LAI, RF, OYI, MN and SF2), C (C2220, 92BR025.8), D (NDK, Z2Z3, ELI, 84ZR085.1, 94UG114.1), F (93BR020.1), and H (90CF056.1) as well as A/C (ZAM184, 92RW009.6), A/G (92NG083.2, 92NG003.1, Z321, IBNG), A/D (MAL), and A/E (93TH253.3, CM240, 90CF402.1) and B/F (93BR029.4) recombinants (SIVcpzGAB was included as an outgroup). All sites with a gap in any of the sequences were removed from the alignment to ensure that all comparisons were made across the same sites. The percent nucleotide sequence diversity between 94CY032.3 and selected other viruses was then calculated for sequence pairs by moving a window of 400 bp in steps of 10 bp along the genome.

Fig. 9 depicts five such distance plots which illustrate the extent of sequence divergence of 94CY032.3 from representatives of subtypes A (92UG037.1), B (LAI), C (C2220), D (ELI) and G/(A) (92NG083.2). The analysis yielded a set of distance curves with very similar (and for the most part superimposable) diversity profiles, suggesting that 94CY032.3 was roughly equidistant from the other subtypes in most regions of its genome (the same results were also obtained when 94CY032.3 was compared to representatives of subtypes A/E, F, and H; data not shown). However, careful inspection of the graphs revealed several small areas of disproportionate sequence similarity involving two of the five reference sequences. For example, at the 3' end of *gag* and the 3' end of *pol*, 92NG083.3 dropped below all others, indicating a relative greater similarity of 94CY032.3 to subtype G. Similarly, in the 5' end of *gag, vif,* and the 3' and 5' end of *env*, 92UG037.1 fell below all others, indicating a relative greater similarity of 94CY032.3 to subtype A. Together, these results suggested that 94CY032.3 contained subtype A and G-like segments, in addition to regions that appeared to be equidistant from the other subtypes.

Relative differences in the extent of sequence similarity as determined by diversity plots (18, 79) or other methods of distance measurement (75) are not always an indicator of recombination, but can reflect variations in the evolutionary rates of the lineages compared. To determine whether 94CY032.3 was truly mosaic,

an exploratory tree analysis was then performed to look for significantly discordant phylogenetic positions for different parts of its genome (Fig. 10). Using the same multiple genome alignment described above, but excluding all known recombinants (except 92NG083.3 and 92NG003.1), unrooted trees were constructed for overlapping fragments of 400 bp, moved in 10 bp increments along the alignment (for subtypes B and D only three representatives were included). Inspection of the resulting topologies revealed that 94CY032.3 changed its phylogenetic position a total of ten times, alternating between subtype A (Fig. 10A, E, G and J; panels 201-600, 4241-4640, 5071-5470 and 6821-7220), subtype G (Fig. 10B, D and H; panels 1101-1500, 3841-4240 and 5471-5870), and an independent position (Fig. 10C, E, I and K; panels 1751-2150, 4641-5040, 5901-6300 and 7901-8300) that was very similar to the one observed in the C2-V3 region (all discordant positions were supported by significant bootstrap values). Since the latter has served as the basis for subtype I definition, it is most parsimonious to assume that all independently grouping segments in 94CY032.3 are of a common origin and thus represent "subtype I". 94CY032.3 thus appears to be comprised of sequences belonging to at least three different (group M) subtypes.

To map the boundaries of the putative A, G and I segments, boostrap plot analyses were performed as previously described (18, 57, 79), plotting the magnitude of the bootstrap values that supported the clustering of 94CY032.3 with 92UG037.1 (subtype A), as well as that of 94CY032.3 with 92NG083.2 ("subtype G"). The results of these analyses allowed us to tentatively map the location and boundaries of the various subtype A an G segments along the 94CY032.3 genome (Fig. 11). Bearing in mind the window size of 400 nucleotides and considering only peaks of significant bootstrap values (>80%), we identified two A/G cross-overs around 1200 and 5600, and one G/A cross-over around 4100. The bootstrap plots also outlined regions with no peaks (or peaks below 80%), which coincided with segments that clustered independently (i.e., in subtype I) in the exploratory tree analysis. Delineating the boundaries of these regions suggested five additional breakpoints: G/I at 1500, I/G at 3800, G/I at 6000, I/A at 6900, and A/I at 7200. Because full length non-mosaic reference sequences for the parental lineages (G and I) were not available, most of the breakpoints could not be mapped with certainty (the

A/G breakpoints at 1200 and 5600 were confirmed by informative site analysis; data not shown). Also, the recombinant nature of 92NG083.2 prohibited reliable breakpoint analysis between 4200 and 4800 (32, 79; highlighted in Fig. 11).

INS H28

~~To map potential recombination breakpoints in this remaining region,~~ four recently reported, partial but non-mosaic subtype G sequences from Mali which spanned the *vif/vpr* region and thus bridged the "subtype A gap" of 92NG083.2 were used (77). A set of distance plots that compare 94CY032.3 to one of these newly derived G sequences (95ML045) as well as representatives of subtype A (U455), B (MN), and D (ELI), respectively, were constructed (data not shown). Consistent with the results from the exploratory tree analysis (Fig. 4), 94CY032.3 was disproportionately more closely related to U455 in the 5' and 3' thirds of this fragment, suggesting the presence of subtype A-like segments. However, in the middle of the fragment, 94CY032.3 was clearly equidistant from U455 and the other subtypes, suggesting an independent position (diversity plots were generated for a window of 300 bp moved in increments of 10 bp). Thus, noting the points at which the "A" distance increased and decreased relative to the other distances allowed us to tentatively map the two remaining breakpoints, one at 4650 and the other at 5000. Trees constructed from sequences surrounding these two breakpoints (Fig. 12) confirmed that 94CY032.3 switched position from subtype A (Fig. 12; panel 4255-4650) to subtype I (panel 4651-5000), and back to subtype A (5001-5300; ~~note, that the new subtype G sequences only cover the region between 4255 and 5300)~~.

There are a total of 10 recombination breakpoints between the 5' end of *gag* and the 3' end of *nef* in the genome structure of the 94CY032.3. However, the discordant subtype assignments of *gag* and *nef* regions necessitate at least one more breakpoint in the viral LTR or the *gag* leader sequence (LTR sequences were not separately analyzed for mosaicism). Given this extent of mosaic complexity, 94CY032.3 is likely the result multiple sucessive recombination events.

Having identified several fragments of subtype I in 94CY032.3, evidence for its presence in other (full length) recombinants from the database was examined. (Data not shown) Two known mosaics MAL (53, 76) and Z321 (78) were of particular interest, because previous analyses had indicated that these viruses contain regions of uncertain subtype assignment (53, 82, 83). For example, MAL has

long been known to represent a mosaic of subtypes A and D, but also contains a sizable *pol* fragment that has defied previous subtype classification (53, 83). Similarly, Z321 is a known mosaic of subtypes A and G (78), but a recent re-analysis of its recombination breakpoints identified regions that could not be assigned to any known subtype (82, 83). To determine whether any of these regions represented subtype I, distance plot analysis was performed, comparing the diversity profiles of MAL and Z321 with those for representatives of other subtypes. Looking for dips in the curves as an indication of relatively greater sequence similarity, one in the *pol* region of MAL and another in the *vif/vpr* region of Z321 were found to coincide with previously unclassified segments of their genomes (indicated as white boxes). Phylogenetic tree analysis confirmed that these regions were indeed of subtype I origin, since MAL and Z321 clustered significantly with the subtype I domains of 94CY032.3. Interestingly, subtype I did not account for all of the unclassifiable regions in MAL and Z321 (82, 83). It thus remains unclear whether these represent still other, as yet unidentified, subtypes or regions of multiple breakpoints that cannot be mapped using current methods.

The above results demonstrate that a strain of HIV-1, proposed in 1995 as a prototypic "subtype I" isolate (29), represents a complex mosaic comprised of subtypes A, G and I, respectively. In addition, two of the oldest known isolates from Africa, MAL (isolated in 1984) (76) and Z321 (isolated in 1976) (80, 84), are shown to contain short segments of sequence closely related to the subtype I domains of 94CY032.3. These findings support the following conclusions: (i) although initially detected in Cyprus, subtype I must have existed in Africa as early as 1976; it is unknown whether full length non-mosaic representatives of subtype I still exist (but have not yet been sampled), or whether this subtype (like subtype E) is represented only by fragments in present day recombinants; (ii) the ancestry of 94CY032.3 must have involved multiple successive recombination events; it remains unclear whether this occurred in Africa and/or in Cyprus, where a number of different subtypes have also been documented (29); (iii) subtype I, along with subtypes A and G, must have diverged substantially earlier than the 1970s in order to be detectable as distinct segments in the Z321 genome; this is consistent with the recent molecular characterization of a virus from 1959 which in phylogenetic analyses appears to have

postdated the group M radiation (85); (iv) finally, the finding of subtype I in several different recombinants, including one from an intravenous drug user (29), suggests that this subtype may be more widespread than previously thought, at least in the form of mosaic genome fragments. It will be interesting to screen additional viruses from drug user populations and their contacts in Cyprus and Greece to determine the current prevalence and geographic distribution of subtype I containing viruses.

REFERENCES

1. Abimiku, A. G., *et al.*, 1994. Subgroup G HIV type 1 isolates from Nigeria. *AIDS Res. Hum. Retroviruses* **10**:1581-1583.

2. Ausubel, F. M., *et al.*, 1987. Current protocols in molecular biology. John Wiley & Sons, New York.

3. Betts, M. R., *et al.*, 1997. Cross-clade HIV-specific cytotoxic T-lymphocyte responses in HIV-infected Zambians. *J. Virol.*, **71**:8908-8911.

4. Bobkov, A., *et al.*, 1996. Complex mosaic structure of the partial envelope sequence from a Gambian HIV Type 1 Isolate. *AIDS Res. Hum. Retroviruses* **12**:169-171.

5. Brodine, S. K., J. R. Mascola, and F. E. McCutchan. 1997. Genotypic variation and molecular epidemiology of HIV. *Infect .Med ., 14:739-748* .

6. Cao, H., *et al.*, 1997. Cytotoxic T-lymphocyte cross-reactivity among different human immunodeficiency virus type 1 clades: Implications for vaccine development. *J. Virol.* **71**:8615-8623.

7. Carr, J. K., *et al.*, 1996. Full length sequence and mosaic structure of a human immunodeficiency virus type 1 isolate from Thailand. *J. Virol.* **70**:5935-5943.

8. Cornelissen, M., *et al.*, 1996. Human immunodeficiency virus type 1 subtypes defined by env show high frequency of recombinant gag genes. *J. Virol.* **70**:8209-8212.

9. Dittmar, M. T., *et al.*, 1997. Langerhans cell tropism of human immunodeficiency virus type 1 subtype A through F isolates derived from different transmission groups. *J. Virol.* **71**:8008-8013.

10. Dolin, R. 1995. Human studies in the development of human immunodeficiency virus vaccines. *J. Infect. Dis.* **172**:1175-1183.

58

11. Esparaza, J., S. Osmanov, and W. Heyward. 1995. HIV preventive vaccines. *Drugs* **50**:792-804.

12. Faulkner, D. M., and J. Jurka. 1988. Multiple aligned sequence editor (MASE). *Trends Biochem. Sci.* **13**:321-322.

13. Felsenstein, J. 1985. Confidence limits on phylogenies: an approach using the boostrap. *Evolution* **39**:783-791.

14. Felsenstein, J. 1992. PHYLIP (Phylogeny Inference Package), 3.5c ed. Department of Genetics, University of Washington, Seattle, Washington.

15. Ferrari, G., *et al.*, 1997. Clade B-based HIV-1 vaccines elicit cross-clade cytotoxic T lymphocyte reactivities in uninfected volunteers. *Proc. Natl. Acad. Sci. USA* **94**:1396-1401.

16. Gao, F. and B. H. Hahn, unpublished.

17. Gao, F., *et al.*, 1994. Genetic variation of HIV type 1 in four World Health Organization-sponsored vaccine evaluation sites: generation of functional envelope (glycoprotein 160) clones representative of sequence subtypes A, B, C, and E. *AIDS Res. Hum. Retroviruses* **10**:1359-1368.

18. Gao, F., *et al.*, 1996. The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin. *J. Virol.* **70**:7013-7029.

19. Gao, F., *et al.*, 1996. Molecular cloning and analysis of functional envelope genes from HIV-1 sequence subtypes A through G. *J. Virol.* **70**:1651-1667.

20. Ghosh, S. K., *et al.*, 1993. A molecular clone of HIV-1 tropic and cytopathic for human and chimpanzee lymphocytes. *Virology* **194**:858-864.

21. Graham, B. S., and P. F. Wright. 1995. Candidate AIDS Vaccines. *N. Engl. J. Med.* **333**:1331-1339.

22. Hahn, B. H., *et al.*, 1984. Molecular cloning and characterization of the HTLV-III virus associated with AIDS. *Nature* **312**:166-169.

23. Hahn, B. H., D. L. Robertson, and P. M. Sharp. 1995. Intersubtype recombination in HIV-1 and HIV-2, p. III-22 - III-29. In G. Myers and B. Korber and S. Wain-Hobson and K.-T. Jeang and L. E. Henderson and G. N. Pavlakis (ed.), Human retroviruses and AIDS1995: A compilation and analysis of nucleic acid and amino acid sequences. Los Alamos National Laboratory, Los Alamos, N. M.

24. Hu, D. J., *et al.*, 1996. The emerging diversity of HIV: the importance of global surveillance for diagnostics, research and prevention. *JAMA* **275**:210-216.

25. Kalish, M. L., *et al.*, 1995. The evolving molecular epidemiology of HIV-1 envelope subtypes in injecting drug users in Bangkok, Thailand: implications for HIV vaccine trials. *AIDS* **9**:851-857.

26. Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* **16**:111-120.

27. Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, U.K.

28. Korber, B. T. M., *et al.*, . 1994. The World Health Organization Global Programme on AIDS Proposal for Standardization of HIV Sequence Nomenclature. *AIDS Res. Hum. Retroviruses* **10**:1355-1358.

29. Kostrikis, L. G., *et al.*, 1995. Genetic analysis of human immunodeficiency virus type 1 strains from patients in Cyprus: Identification of a new subtype designated subtype I. *J. Virol.* **69**:6122-6130.

30. Leitner, T., and J. Albert. 1995. A new genetic subtype of HIV-1, p. III-147 - III-150. In G. Myers and B. Korber and B. H. Hahn and K.-T. Jeang and J. W. Mellors and F. E. McCutchan and L. E. Henderson and G. N. Pavlakis (ed.),

Human Retroviruses and AIDS 1995. Theoretical Biology and Biophysics, Los Alamos.

31. Leitner, T., *et al.*, 1995. Biological and molecular characterization of subtype D, G, and A/D recombinant HIV-1 transmissions in Sweden. *Virology* **209**:136-146.

32. Leitner, T., B.T.M. Korber, D.L. Robertson, F. Gao, and B.H.Hahn. 1997. Updated Proposal of Reference Sequences of HIV-1 Genetic Subtypes. *In* B. Korber, B. Foley, C. Kuiken, T. Leitner, F. McCutchan, J. W. Mellors and B. H. Hahn (ed), Human Retroviruses and AIDS 1997: a complication and analysis of nucleic acid and amino acid sequence. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico.

33. Loussert-Ajaka, I., *et al.*, 1995. Variability of human immunodeficiency virus type 1 group O strains isolate from Cameroonian patients living in France. *J. Virol.* **69**:5640-5649.

34. Louwagie, J., *et al.*, 1993. Phylogenetic analysis of gag genes from 70 international HIV-1 isolates provides evidence for multiple genotypes. *AIDS* 7:769-780.

35. Louwagie, J., *et al.*, 1995. Genetic diversity of the envelope glycoprotein from human immunodeficiency virus type 1 isolates of African origin. *J. Virol.* **69**:263-271.

36. Maniatis, T., E. F. Fritsch, and J. Sambrook. 1982. Molecular cloning: a laboratory manual, p. 269-295. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

37. Martin-Gallardo, A., J. Lamerdin, and A. Carrano. 1994. Shotgun sequencing, p. 37-41. In M. D. Adams and C. Fields and J. C. Venter (ed.), Automated DNA sequencing and analysis. Academic Press, London.

38. Mascola, J. R., *et al.*, 1996. Human immunodeficiency virus type 1 neutralizing antibody serotyping using serum pools and an infectivity reduction assay. *AIDS Res. Hum. Retroviruses* **12**:1319-1328.

39. McCutchan, F. E., *et al.*, 1992. Genetic variants of HIV-1 in Thailand, *AIDS Res. Hum. Retroviruses* **8**:1887-1895.

40. McCutchan, F. E., M. O. Salminen, J. K. Carr, and D. S. Burke. 1996. HIV-1 genetic diversity. *AIDS* **10**(suppl 3):S13-20.

41. Moore, J. P., *et al.*, 1996. Inter- and intra- subtype neutralization of human immunodeficiency virus type 1: the genetic subtypes do not correspond to neutralization serotypes but partially correspond to gp120 antigenic serotypes. *J. Virol.* **70**:427-444.

42. Moore, J., and A. Trkola. 1997. HIV type 1 coreceptors, neutralization serotypes, and vaccine development. *AIDS Res. Hum. Retroviruses* **13**:733-736.

43. Murphy, E., *et al.*, 1993. Diversity of V3 region sequences of human immunodeficiency viruses type 1 from the Central African Republic. *AIDS Res. Hum. Retroviruses* **9**:997-1007.

44. Myers, G., *et al.*, 1992. Human retroviruses and AIDS: a compilation and analysis of nucleic acid and amino acid sequence. Theoretical Biology and Biophysics Group,Los Alamos National Laboratory, Los Alamos, New Mexico.

45. Myers, G., *et al.*, 1996. Human retroviruses and AIDS: A compilation and analysis of nucleic acid amino acid sequences. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, New Mexico.

46. Nyambi, P. N., *et al.*, 1996. Multivariate analysis of human immunodeficiency virus type 1 neutralization data. *J. Virol.* **70**:445-458.

47. Ou, C.-Y., *et al.*, 1993. Independent introductions of two major HIV-1 genotypes into distinct high-risk populations in Thailand. *Lancet* **341**:1171-1174.

48. Peden, K., M. Emerman, and L. Montagnier. 1997. Changes in growth properties on passage in tissue culture of viruses derived from infectious molecular clones of HIV-1LAI, HIV-1MAL, and HIV-1ELI. *Virology* **185**:661-672.

49. Perrière, G. and Gouy, M. 1996. WWW-Query: An on-line retrieval system for biological sequence banks. Biochimie **78**: 364-369.

50. Pope, M., *et al.*, 1997. HIV-1 strains from subtypes B and E replicate in cutaneous dendritic cell-T cell mixtures without displaying subtype-specific tropism. *J. Virol.* **71**:8001-8007.

51. Pope, M., *et al.*, 1997. Different subtypes of HIV-1 and cutaneous dendritic cells. *Science* **278**:786-787.

52. Robertson, *et al.*, 1995. Recombination in HIV-1. *Nature* **374**:124-126.

53. Robertson, D. L., B. H. Hahn, and P. M. Sharp. 1995. Recombination in AIDS viruses. *J. Mol. Evol.* **40**:249-259.

54. Sabino, E. C., *et al.*, 1994. Identification of human immunodeficiency virus type 1 envelope genes recombinant between subtypes B and F in two epidemiologically linked individuals from Brazil. *J. Virol.* **68**:6340-6346.

55. Saitou, N., and M. Nei. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**:406-425.

56. Salminen, M. O., *et al.*, 1995. Recovery of virtually full length HIV-1 provirus of diverse subtypes from primary virus cultures using the polymerase chain reaction. *Virology* **213**:80-86.

57. Salminen, M. O., J. K. Carr, D. S. Burke, and F. E. McCutchan. 1995. Identification of breakpoints in intergenotypic recombinants of HIV-1 by bootscanning. *AIDS Res. Hum. Retroviruses* **11**:1423-1425.

58. Salminen, M. O., J. K. Carr, D. S. Burke, and F. E. McCutchan. 1995. Genotyping of HIV-1, p. III-30 - III-34. In G. Myers and B. Korber and S. Wain-Hobson and

K.-T. Jeang and L. E. Henderson and G. N. Pavlakis (ed.), Human retroviruses and AIDS 1995: A compilation and analysis of nucleic acid and amino acid sequences. Los Alamos National Laboratory, Los Alamos.

59. Salminen, M. O., *et al.*, 1996. Full length sequence of an ethiopian human immunodeficiency virus type 1 (HIV-1) isolate of genetic subtype C. *AIDS Res. Hum. Retroviruses* **12**:1329-1339.

60. Salminen, M. O., *et al.*, 1997. Evolution and probable transmission of intersubtype recombinant human immunodeficiency virus type 1 in a Zambian couple. *J. Virol.* **71**:2647-2655.

61. Sharp, P. M., D. L. Robertson, F. Gao, and B. H. Hahn. 1994. Origins and diversity of human immunodeficiency viruses. *AIDS* **8**:S27-S42.

62. Sharp, P. M., D. L. Robertson, and B. H. Hahn. 1995. Cross-species transmission and recombination of AIDS viruses. *Phil. Trans. R. Soc. London (Ser. B)* **349**: 41-47.

63. Siepel, A. C., and B. T. Korber. 1995. Scanning the database for recombinant HIV-1 genomes, p. III-35 - III-60. In G. Myers and B. Korber and S. Wain-Hobson and K.-T. Jeang and L. E. Henderson and G. N. Pavlakis (ed.), Human retroviruses and AIDS 1995: A compilation and analysis of nucleic acid and amino acid sequences. Los Alamos National Laboratory, Los Alamos, N. M.

64. Soto-Ramirez, L. E., *et al.*, 1996. HIV-1 langerhans' cell tropism associated with heterosexual transmission of HIV. *Science* **271**:1291-1293.

65. Spire, B., *et al.*, 1989. Nucleotide sequence of HIV1-NDK: a highly cytopathic strain of the human immunodeficiency virus. *Gene* **81**:275-284.

66. Takehisa, J., *et al.*, 1997. Phylogenetic analysis of human immunodeficiency virus 1 in Ghana. *Acta. Virologia* **41**:51-54.

67. Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W - improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**:4673-4680.

68. Weber, J., *et al.*, 1996. Neutralization serotypes of human immunodeficiency virus type 1 field isolates are not predicted by genetic subtype. *J. Virol.* **70**:7827-7832.

69. Weniger, B. G., *et al.*, 1991. The epidemiology of HIV infection in AIDS in Thailand. *AIDS* **5**(suppl. 2):S71-S85.

70. Weniger, B. G., *et al.*, 1994. The molecular epidemiology of HIV in Asia. *AIDS* **8**(suppl. 2):S13-S28.

71. Wieland, U., *et al.*, 1997. Diversity of the vif gene of human immunodeficiency virus type 1 in Uganda. *J. of Gen. Virol.* **78**:393-400.

72. World Health Organization Network for HIV Isolation and Characterization. 1994. HIV-1 variation in WHO-sponsored vaccine-evaluation sites: Genetic screening, sequence analysis and preliminary biological characterization of selected viral strains. *AIDS. Res. Hum. Retroviruses* **10**:1327-1344.

73. Zhang, L., *et al.*, 1996. HIV-1 subtype and second-receptor use. *Nature (London)* **383**:768.

74. Zhang, L., *et al.*, 1997. HIV-1 subtypes, co-receptor usage, and CCR5 polymorphism. *AIDS Res. Hum. Retroviruses* **13**: 1357-1366

75. Siepel, A.C., *et al.*, A computer program designed to screen rapidly for HIV type 1 intersubtype recombinant sequences. *AIDS Res. Hum. Retrovirus.* **11**: 1413-1416, 1995

76. Alizon, M., S. *et al.*, 1986. Genetic variability of the AIDS virus: nucleotide sequence analysis of two isolates from African patients. *Cell* **46**:63-74.

77. Bibollet-Ruche, F., *et al.*, Genetic characterization of accessory genes from human immunodeficiency virus type 1 subtypes A, C, D, F, G and H from different African countries, in preparation.

78. Choi, D. J., *et al.*, 1997. HIV type 1 isolate Z321, the strain used to make a therapeutic HIV type 1 immunogen, is intersubtype recombinant. *AIDS Res. Hum. Retroviruses* **13**:357-361.

79. Gao, F., *et al.*, 1998. A comprehensive panel of near full-length clones and reference sequences for non-subtype B isolates of human immunodeficiency virus type 1. *J. Virol*, in press.

80. Getchell, J. P., *et al.*, 1987. Human immunodeficiency virus isolated from a serum sample collected in 1976 in Central Africa. *J. Infect. Dis.* **156**:833-837.

81. B. Korber, *et al.*, 1997. Human retroviruses and AIDS: A compilation and analysis of nucleic acid and amino acid sequences. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, N.M.

82. Robertson, D. L., F. Gao, and B. H. Hahn. The analysis of complete HIV-1 intersubtype hybrid genomes, in preparation.

83. Robertson, D. L., *et al.*, 1997. Intersubtype Recombinant HIV-1 Sequences. pp. III 25-III 30, *In* B. Korber, B. Foley, C. Kuiken, T. Leitner, F. McCutchan, J. W. Mellors, and B. H. Hahn (ed.), Human retroviruses and AIDS 1997: A compilation and analysis of nucleic acid and amino acid sequences. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, N.M.

84. Srinivasan, A., D. *et al.*, 1989. Molecular characterization of HIV-1 isolated from a serum collected in 1976: Nucleotide sequence comparison of recent isolates and generation of hybrid HIV. *AIDS Res. Hum. Retroviruses* **5**:121-129.

85. Zhu, T., *et al.*, 1998. An African HIV-1 sequence from 1959 and implications for the origin of the epidemic. *Nature* **391**:594-597.

86. Southern, E.M. 1975. *J. Mol. Biol.*, **98**:503-517.

87. Kafatos, F.C. *et al.* 1979. *Nucleic Acids Res.*, 7:1541-1522

66

88. Agarwal *et al.* 1972, *Angew. Chem. Int. Ed. Engl.* **11**:451.

89. Baeucage *et al.* 1981, *Tetrahedron Letters* **22**:1859-1862. Automated diethylphosphoramidite method.

90. Hsiung *et al.* 1979. *Nucleic Acids Res* **6**:1371

91. See, e.g., Anderson, *et al.* 1996. *Antimicrob. Agents Chemother.*, **40**:2004-2011; Azad, *et al.* 1995. *Antiviral Res.*, **28**:101-111; Azad, *et al.* 1993. *Antimicrob. Agents Chemother.*, **37**:1945-1954; Leeds, *et al.* 1997. *Drug. Metab. Dispos.*, **25**:921-926; and references therein. See also, Cook, P.D., 1993. Monomers for preparation of oligonucleotides having chiral phosphorus linkages. US Patent 5,212,295 (re: general method of making DNA analogs, including phosphorothioates, thioesters, etc.); and Iyer *et al.* 1990 *J. Org. Chem.* **55**:4693-4699 (re: synthetic method for making phosphorothioate oligos).

92. See, e.g., Nielsen, *et al.*, WO 98/03542; Hyrup and Nielsen 1996. *Bioorg. Med. Chem.* **4**:5-23; and Nielsen, *et al.* 1991. *Science* **254**:1497-1500; and references therein.

93. Sambrook, J. *et al.* 1989. In "Molecular Cloning, A Laboratory Manual", Cold Spring Harbor Press, Plainview, New York.

94. Alwine, J.C., *et al.* 1977. *Proc. Natl. Acad. Sci.*, **74**:5350-5354.

95. Hollander, M.C. *et al.* 1990. *Biotechniques*; **9**:174-179

96. Watson, J.D., *et al.* 1992. In "Recombinant DNA" Second Edition, W.H. Freeman and Company, New York.

97. See, e.g., Naldini, N., *et al.*, "In Vivo Gene Delivery and Stable Transduction of Nondividing Cells by a Lentiviral Vector", *Science*, **272**:263-267 (1996); Srinivasakumar, N., *et al.*, "The Effect of Viral Regulatory Protein Expression on Gene Delivery by Human Immunodeficiency Virus Type 1 Vectors Produced in Stable Packaging Cell Lines", *J. Virol.*, **71**:5841-5848 (Aug. 1997); Zufferey, R., *et al.*, "Multiply Attenuated Lentiviral Vector Achieves Efficient Gene-Delivery In Vivo", *Nature Biotechnology*, **15**:871-875 (Sept. 1997); and Kim, V.N., *et al.*,

"Minimal Requirement for a Lentivirus Vector Based on Human Immunodeficiency Virus Type 1", *J. Virol.*, **72**:811-816 (Jan. 1998).

98. *See, e.g.*, Schwartz *et al.*, *J. Virol.*, **66**:7176-7182 (1992); International Publication No. WO 93/20212 (1993); Schneider, R., *et al.*, "Inactivation of the human immunodeficiency virus type 1 inhibitory elements allows Rev-independent expression of Gag and Gag/protease and particle formation," *J. Virol.* , **71**:4892-4903 (1997) concerning the identification and mutation of inhibitory and instability regions using multiple point mutations within HIV-1 *gag*, *protease* and *pol* coding regions to reduce the effects of these regions and increase expression of the encoded polypeptide.

99. Oellerich, M. 1984. J. Clin. Chem. Clin. BioChem 22:895-904

100. Lu S., *et al.* Simian immunodeficiency virus DNA vaccine trial in macaques. *J. Virol.* 1996;70:3978-91.

101. Haynes JR, *et al.*, Accell particle-mediated DNA immunization elicits humoral, cytotoxic and protective responses. *AIDS Res. Hum. Retroviruses* 1994; 10 (suppl 2): S43-45

102. Okuda, K, *et al.* Induction of potent humoral and cell-mediated immune responses following direct injection of DNA encoding the HIV type 1 Env and Rev gene products. *AIDS Res. Hum. Retroviruses* 1995;11:933-43

103. Wang B., *et al.* Induction of humoral and cellular immune responses to the human immunodeficiency type 1 virus in non-human primates by in vivo DNA inoculation. *J. Virol.* 1995; 21:102-12

104. Boyer JD, *et al.* In vivo protective anti-HIV immune responses in non-human primates through DNA immunization. *J. Med. Primatol.* 1996; 25-242-50

105. MacGregor *et al.*, *J. Infect. Dis.* 178:92-100 (1998)

106. Donnelly *et al.*, *Annu. Rev. Immunol.* 15:617-648 (1997)

107. Ulmer *et al.*, *Science* 259:1745-1749 (1993)

108. Winzeler *et al.*, *Science* 281:1194-1197 (1998)

68

Modifications of the above described invention that are obvious to those of skill in the fields of genetic engineering, immunology, virology, protein chemistry, medicine, and related fields are intended to be within the scope of the following claims.

All of the references cited herein above are hereby incorporated by reference.

INS ~~B1~~ B1

69